Identifiability and Efficient Estimation With Nonignorable Response Mechanism

KENJI BEPPU

MARCH 2025

Identifiability and Efficient Estimation With Nonignorable Response Mechanism

A dissertation submitted to
THE GRADUATE SCHOOL OF ENGINEERING SCIENCE
OSAKA UNIVERSITY
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY IN ENGINEERING

BY
KENJI BEPPU

MARCH 2025

Abstract

Handling missing data inheres significant challenges in statistical analysis due to the potential for substantial bias and efficiency loss, particularly when the response mechanism depends on unobserved variables, which is referred to as "nonignorable". A key difficulty in analyzing nonignorable missing data is guaranteeing identifiability. A model that lacks identifiability cannot be estimated theoretically, and inferences made from such a model tend to be unstable. To address these issues, this study develops methodologies to ensure model identifiability in the presence of nonignorable nonresponse.

First, this research explores the connection between model identifiability and the completeness condition under instrumental variable settings. It introduces a novel sufficient condition based on the monotone-likelihood property, providing a practical alternative when establishing the completeness condition is difficult. The second proposed approach focuses on modeling the joint distribution of observed and unobserved variables using a logistic response mechanism and generalized linear outcome models. A key contribution of this work is the derivation of sufficient conditions for model identifiability without relying on instrumental variables, which are often assumed to ensure identifiability but are typically challenging to identify in practice. Numerical simulations and real data analyses illustrate the practical effectiveness of these methods, highlighting their potential for addressing nonignorable missing data in real-world applications.

The discussion further transitions to survey sampling, a field where missing data is commonly problematic. In particular, nonresponse occurred with nonnegligible probability after probability sampling presents a pervasive issue in survey sampling, often requiring adjustments to simultaneously address both sampling and selection biases. This study examines strategies for reducing bias and optimizing the use of available information, not only in nonresponse scenarios but also in data integration settings, where summary statistics from external data sources are available. We reframe these challenges within a two-step monotone missing data framework: the first stage involves missingness due to sampling, while the second focuses on nonresponse. To enhance robustness against model misspecification within this framework, we extend the concept of double robustness to multiple robustness by introducing a two-step empirical likelihood method. This method efficiently leverages empirical weights to improve estimation accuracy. In addition, we illustrate how the identifiability conditions established in the first part can be applied not only to handling missing data but also to integrating non-probability samples in survey sampling. We can accomplish this by interpreting the pairing of probability and non-probability samples within the broader context of the incomplete data framework.

Contents

1 Introduction

There has been a rapidly growing movement to utilize all available data, even when such data explicitly or implicitly contain missing values, in fields such as causal inference (Imbens and Rubin, 2015) and data integration (Hu, Wang, Li and Miao, 2022; Yang and Kim, 2020). For such datasets, handling missing data is a critical issue in many research areas because improper analysis can lead to misleading or erroneous results. For example, "The Strengthening Analytical Thinking for Observational Studies Initiative", a large-scale collaborative research project involving over 100 experts from diverse fields of biostatistics research, identified nine crucial topics in observational studies (Sauerbrei, Abrahamowicz, Altman, le Cessie, Carpenter and initiative, 2014); missing data problems are one of the nine critical topics. The complexity of handling missing data is further heightened when the response mechanism is nonignorable or missing not at random (MNAR), where the missingness is caused by the value that would have been observed (Ibrahim, Chen, Lipsitz and Herring, 2005; Kim and Yu, 2011; Little and Rubin, 2019; Wang, Shao and Kim, 2014).

The nonignorable response mechanism is frequently encountered in practical applications and has been extensively studied in various fields such as RNA-sequencing data analysis (Hicks, Townes, Teng and Irizarry, 2018), school psychology research (Baraldi and Enders, 2010), cost-effectiveness analysis (Leurent, Gomes, Faria, Morris, Grieve and Carpenter, 2018), counselling psychology research (Parent, 2013), clinical trials (Hazewinkel, Bowden, Wade, Palmer, Wiles and Tilling, 2022), and information systems research (Peng, Hahn and Huang, 2023). For example, Peng et al. (2023) reported how information systems researchers deal with missing values and presented six scenarios with R&D data from the Compustat database and online reviews in the field of information systems.

In recent years, analysis of missing data under the missing at random (MAR) assumption (Little and Rubin, 2019) has become increasingly well-established (Kim and Shao, 2021; Robins, Rotnitzky and Zhao, 1994*a*). Although model identifiability is a fundamental requirement for constructing asymptotic theory, relaxing the MAR assumption makes statistical inference drastically difficult, especially in model identification (Miao, Ding and Geng, 2016). Estimation with unidentifiable models may provide multiple solutions that have exactly the same model fitting. Several researchers have considered giving sufficient conditions for the model identification under the MNAR assumption.

Constructing observed likelihood consists of two distributions: response mechanism response mechanism and outcome distribution outcome distribution (Kim and Shao, 2021). Miao et al. (2016) considered identification condition with Logistic, Probit, and Robit (cumulative distribution function of t-distribution) models for response mechanism and normal and t (mixture) distributions for outcome distribution. Cui, Guo and Yang (2017) assumed Logistic, Probit, and cLog-log models for response mechanism and the generalized linear models for outcome distribution. These studies depend heavily on the model specification of both response mechanism and outcome distribution. Wang et al. (2014) introduced a covariate called instrument or shadow variable and demonstrated that the use of the instrument could considerably relax conditions on response mechanism and outcome distribution. For example, outcome distribution requires only the monotone-likelihood property, which includes a variety of models, such as the generalized linear model. Tang, Little and Raghunathan (2003) and Miao and Tchetgen (2018) derived conditions for model identifiability without postulating any assumptions on response mechanism with the help of the instrument. Miao, Liu, Tchetgen and Geng (2019) further relaxed the assumption under an assumption referred to as the completeness condition on response mechanism (D'Haultfœuille, 2010, 2011). For example, the generalized linear model with continuous covariates satisfies the completeness condition. To the best of our knowledge, this combination of an instrument on response mechanism and completeness on outcome distribution is the most general condition for model identification and has been accepted in numerous studies (Yang, Wang and Ding, 2019; Zhao and Ma, 2022).

Generally, assumptions on outcome distribution rely on the distribution of the complete data, which is untestable from observed data. Instead of modelling the distribution of complete data, numerous recently developed methods have modelled the distribution of observed data, referred to as the respondents' outcome model (Kim and Yu, 2011; Li, Ma and Zhao, 2021; Morikawa and Kim, 2021; Riddles, Kim and Im, 2016; Shetty, Ma and Zhao, 2021). This modelling is advantageous because the observed data are available; consequently, we can select a better model for the candidates by using information criteria based on the observed data, such as the Akaike information criterion (AIC), Bayesian information criterion (BIC), or other variable selection methods, such as the adaptive lasso (Zou, 2006). The application of Bayes' theorem yields an explicit expression of the nonrespondents' outcome model from the assumed respondents' outcome model and response mechanism. This alternative expression for the joint distribution of the outcome variable and response indicator is often called Tukey's representation (Franks, Airoldib and Rubin, 2020).

Even after overcoming the challenges of identifiability, estimation methods for nonignorable nonresponse and similar data structures still have significant potential for further development. An example of a data structure similar to nonignorable nonresponse is informative sampling (Pfeffermann, 1993; Pfeffermann and Sverchkov, 2009) in survey sampling. In survey sampling, probability sampling typically provides sampling weights, the inverse of inclusion probabilities, based on covariates known before sampling the outcome variables. When the outcome variables are conditionally independent of covariates given the sampling weights, the sampling is non-informative. However, if some covariates used to construct sampling weights are unavailable to analysts, this dependence between weights and outcomes leads to informative sampling. Morikawa, Terada and Kim (2022) proposed a semiparametric efficient estimator under the informative sampling framework. However, in practice, outcome variables often contain missing values. To address this nonresponse, Kim and Haziza (2014) and Chen and Haziza (2017) proposed a double/multiple robust estimator, respectively. Additionally, Morikawa, Beppu and Aida (2023) derived an adaptive estimator that achieves the semiparametric efficient bound under nonresponse.

In this paper, the discussion centers on two key aspects: identifiability and estimation under informative sampling. Here are the main contributions of each study:

- Identifiability: under the modeling of the respondents' outcome model, we derive two types of identification conditions. First, we consider an identification problem with an instrument for response mechanism and the respondents' outcome that satisfies the monotone-likelihood ratio property. Note that although our model setup is similar to Wang et al. (2014), we can check the validity of the respondents' outcome with observed data. Second, we derive the nonrespondents' outcome model in the form of a generalised linear model without any instrumental or shadow variables when the observed data can be fitted as a generalised linear model and the response mechanism follows a logistic distribution. To estimate the model parameters, we employ fractional imputation (FI), which is among the most beneficial tools in missing data analysis for solving estimating equations (Im, Cho and Kim, 2018;

Kim, 2011). Detailed FI estimation procedures are introduced along with variance estimation by applying the results of Riddles et al. (2016). Given these points, deriving identification conditions without relying on the existence of instrumental variables is useful and accessible to all fields of research.

- Estimation under informative sampling: to overcome the misspecification of the working model, we extend the adaptive estimator proposed by Morikawa et al. (2023) to multiple robust estimators using empirical likelihood weights. Moreover, we develop optimal estimators for informative sampling scenarios using empirical likelihood weights when external summary statistics are accessible. Building on the work of Chatterjee, Chen, Maas and Carroll (2016), Kundu, Tang and Chatterjee (2019), and Zhang, Deng, Schiffman, Qin and Yu (2020), recent advances in data integration utilize both individual-level internal data and summary statistics such as means and variances of covariates and outcomes from external sources. Additionally, Hu et al. (2022) established the efficiency bound for data integration settings and proposed adaptive estimators that achieve this bound. Inspired by their approach, we derive the efficient multiple robust estimator using external summary statistics in our setting.

The remainder of this paper is organised as follows: In section 2, we overview the identification conditions from previous research. Then, we propose the identification conditions for two scenarios. In section 3, the discussion shifts to informative sampling, where we derive an efficient multiple robust estimator using the empirical likelihood approach. Furthermore, as a related topic to the identifiability conditions derived in section 2, we introduce the integration of non-probability samples under nonignorable participation. Numerical examples including a simulation study and real data applications, are presented in Section 4 and 5. Section 6 summarises the concluding remarks. All technical proofs are presented in the Appendix.

2 Identification in nonignorable model

In this section, we focus on the model identifiability in the nonignorable model. First, we introduce the observed likelihood, explaining its formulation and importance in addressing missing data problems. We then discuss estimation methods based on the observed likelihood and other functions. Next, we define the concept of identification within the context of observed likelihood, establishing a clear framework for evaluating identifiability. Building on this background, we propose novel identification conditions.

2.1 Observed likelihood

Let $\{\boldsymbol{x}_i, y_i, \delta_i\}_{i=1}^n$ be independent and identically distributed samples from a distribution of $(\boldsymbol{x}, y, \delta)$, where $\boldsymbol{x}$ is a fully observed covariate vector, y is an outcome variable subject to missingness, and δ is a response indicator of y being 1(0) if y is observed (missing). We use generic notation $p(\cdot)$ and $p(\cdot \mid \cdot)$ for the marginal density and conditional density, respectively. For example, $p(\boldsymbol{x})$ is the marginal density of $\boldsymbol{x}$, and $p(y \mid \boldsymbol{x})$ is the conditional density of y given $\boldsymbol{x}$. We model the MNAR response mechanism $P(\delta = 1 \mid \boldsymbol{x}, y)$ and consider its identification. The observed likelihood is defined as

$$\prod_{i:\delta_i=1} P(\delta_i = 1 \mid y_i, \boldsymbol{x}_i) p(y_i \mid \boldsymbol{x}_i) \prod_{i:\delta_i=0} \int \{1 - P(\delta_i = 1 \mid y, \boldsymbol{x}_i)\} p(y \mid \boldsymbol{x}_i) dy. \tag{1}$$

We say that this model is identifiable if parameters in (1) are identified, which is equivalent to parameters in $P(\delta = 1 \mid y, \boldsymbol{x}) p(y \mid \boldsymbol{x})$ being identified. This identification condition is essential even for semiparametric models such as an estimator defined by moment conditions (Morikawa and Kim, 2021). However, simple models can be easily unidentifiable. For example, Example 1 in Wang et al. (2014) presented an unidentifiable model when the outcome model is normal, and the response mechanism is a Logistic model.

There is an alternative way to express the relationship between y and $\boldsymbol{x}$. A disadvantage of modeling $p(y \mid \boldsymbol{x})$ is its subjective assumption on the distribution of complete data, not of observed data. In other words, if we made assumptions about $p(y \mid \boldsymbol{x})$ and ensured its identifiability, we could not verify the assumptions using the observed data. By contrast, this issue can be overcome by modeling $p(y \mid \boldsymbol{x}, \delta = 1)$ because $p(y \mid \boldsymbol{x}, \delta = 1)$ is the outcome model for the observed data, and we can check its validity using ordinal information criteria such as AIC and BIC. Therefore, we model $p(y \mid \boldsymbol{x}, \delta = 1)$ and consider the identification condition in the later section. Hereafter, we assume two parametric models $p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma})$ and $P(\delta = 1 \mid \boldsymbol{x}, y; \boldsymbol{\phi})$, where $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ are parameters of the outcome and response models, respectively. Although our method requires two parametric models, the class of identifiable models is very large. For example, it can include semiparametric outcome models for $p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma})$ and general response models $P(\delta = 1 \mid \boldsymbol{x}, y; \boldsymbol{\phi})$ other than Logistic models, as discussed in Example 2.10.

2.2 Estimation methods

We present a procedure of parameter estimation based on parametric models of $p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma})$ and $P(\delta = 1 \mid \boldsymbol{x}, y; \boldsymbol{\phi})$. Let $\hat{\boldsymbol{\gamma}}$ be the maximum likelihood estimator of $\boldsymbol{\gamma}$. The observed likelihood (1) yields to the mean score equation for $\boldsymbol{\phi}$ (Kim and Shao, 2021):

$$\sum_{i=1}^n \left\{ \delta_i \frac{\partial \log \pi(\boldsymbol{x}_i, y_i; \boldsymbol{\phi})}{\boldsymbol{\phi}} - (1 - \delta_i) \frac{\int \partial \pi(\boldsymbol{x}_i, y; \boldsymbol{\phi}) / \partial \boldsymbol{\phi} \cdot p(y \mid \boldsymbol{x}) dy}{\int \{1 - \pi(\boldsymbol{x}_i, y; \boldsymbol{\phi})\} p(y \mid \boldsymbol{x}) dy} \right\} = 0, \tag{2}$$

where $\pi(\boldsymbol{x}, y; \boldsymbol{\phi}) = P(\delta = 1 \mid \boldsymbol{x}, y; \boldsymbol{\phi})$. By using Bayes' formula $p(y \mid \boldsymbol{x}) \propto p(y \mid \boldsymbol{x}, \delta = 1)/\pi(\boldsymbol{x}, y; \boldsymbol{\phi})$, the mean score can be written as

$$\sum_{i=1}^{n} \{\delta_i s_1(\boldsymbol{x}_i, y_i; \boldsymbol{\phi}) + (1 - \delta_i) s_0(\boldsymbol{x}_i; \boldsymbol{\phi})\} = 0,$$

where

$$s_1(\boldsymbol{x}, y; \boldsymbol{\phi}) = \frac{\partial \log \pi(\boldsymbol{x}, y; \boldsymbol{\phi})}{\partial \boldsymbol{\phi}}, \ s_0(\boldsymbol{x}; \boldsymbol{\phi}) = -\frac{\int s_1(\boldsymbol{x}, y; \boldsymbol{\phi}) p(y \mid \boldsymbol{x}, \delta = 1) dy}{\int \{1/\pi(\boldsymbol{x}, y; \boldsymbol{\phi}) - 1\} p(y \mid \boldsymbol{x}, \delta = 1) dy}.$$

To compute the two integrations in $s_0(\cdot)$, we can use the fractional imputation (Kim, 2011). As described in Riddles et al. (2016), the EM algorithm is also applicable.

Many identification conditions have been proposed under nonignorable missing data because identifiability conditions exist for each estimation method. For example, identifiability for maximizing the observed likelihood (1) or pseudo-likelihood that is a certain function. First, we explain the definition of identifiability under the estimation using the observed likelihood in the section 2.3. Second, we overview the previous research for each modeling. Subsequently, we propose new identification conditions.

2.3 Definition of identification in the use of observed likelihood

Recall that the identification condition in (1) is for parameters in $P(\delta = 1 \mid y, \boldsymbol{x}) p(y \mid \boldsymbol{x})$. As seen in Section 2.2, the conditional density $p(y \mid \boldsymbol{x})$ is represented by $p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma})$ and $P(\delta = 1 \mid \boldsymbol{x}, y; \boldsymbol{\alpha}, \boldsymbol{\phi})$ by Bayes' formula. Thus, using the formula, identification with these models changes to parameters in $\varphi(y, \boldsymbol{x}; \boldsymbol{\phi}, \boldsymbol{\gamma})$, where

$$\varphi(y, \boldsymbol{x}; \boldsymbol{\phi}, \boldsymbol{\gamma}) = \frac{p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma})}{\int p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma})/\pi(\boldsymbol{x}, y; \boldsymbol{\phi}) dy}. \tag{3}$$

Strictly speaking, the identification condition is $\varphi(y, \boldsymbol{x}; \boldsymbol{\phi}, \boldsymbol{\gamma}) = \varphi(y, \boldsymbol{x}; \boldsymbol{\phi}', \boldsymbol{\gamma}')$ with probability 1 implies that $(\boldsymbol{\phi}^\top, \boldsymbol{\gamma}^\top) = (\boldsymbol{\phi}'^\top, \boldsymbol{\gamma}'^\top)$. Generally, the integral in the denominator of (3) does not have the closed form, which makes deriving a sufficient condition for the identifiability quite challenging. Morikawa and Kim (2021) identified a combination of Logistic models and normal distributions for response and outcome models has a closed form of the integration and derived a sufficient condition for the model identifiability. Beppu, Choi, Morikawa and Im (2024) extended the model to a case where the outcome model belongs to the exponential family while the response model is still a Logistic model. However, when the response mechanism is general, simple outcome models such as normal distribution can be unidentifiable.

Example 2.1. Suppose that the respondents' outcome model is $y \mid (\delta = 1, x) \sim N(\gamma_0 + \gamma_1 x, 1)$, and the response model is $P(\delta = 1 \mid x, y) = \Psi(\alpha_0 + \alpha_1 x + \beta y)$, where Ψ is a known distribution function such that the integration in (3) exists and the covariate x is one-dimensional vector in this example. Then, this model is unidentifiable. For example, different parametrization $(\alpha_0, \alpha_1, \beta, \gamma_0, \gamma_1) = (0, 1, 1, 0, 1)$, $(\alpha_0', \alpha_1', \beta', \gamma_0', \gamma_1') = (0, 3, -1, 0, 1)$ yields the same value of the observed likelihood.

2.4 Previous research on identifiability

In this section, we review the identification conditions proposed in previous research. Because identifiability depends on the estimation method, we first focus on identifiability related to the maximization of the observed likelihood. Subsequently, we introduce the identifiability for the generalized method of moments and stably estimation proposed by Li, Qin and Liu (2023) as alternative methods to observed likelihood maximization. Note that there are various estimation methods besides maximizing the observed likelihood, each with its own identification conditions. For example, Tang et al. (2003), Liu, Li and Qin (2022) and the conditional likelihood approaches proposed in section 6.2 of Kim and Shao (2021).

2.4.1 Identifiability of the observed likelihood

Lemma 2.2. *(Identification condition proposed by Wang et al. (2014)) The observed likelihood (1) is identifiable under the following conditions:*

1. *The covariate $\boldsymbol{x}$ has two components, $\boldsymbol{x} = (\boldsymbol{u}^\top, \boldsymbol{z}^\top)^\top$, such that*

$$\pi(\boldsymbol{x}, y; \boldsymbol{\phi}) = P(\delta = 1 \mid y, \boldsymbol{u}; \boldsymbol{\phi}) = F(\alpha_{\boldsymbol{u}} + \beta_{\boldsymbol{u}} y), \tag{4}$$

   where $\alpha_{\boldsymbol{u}}$ and $\beta_{\boldsymbol{u}}$ are unknown parameters depending on only $\boldsymbol{u}$, not $\boldsymbol{z}$, the function $F : \mathbb{R} \to (0, 1]$ is a known strict monotone and twice differentiable, and for any given $\boldsymbol{u}$, there exists two values of $\boldsymbol{z}$, $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, such that $p(y \mid \boldsymbol{u}, \boldsymbol{z}_1) \neq p(y \mid \boldsymbol{u}, \boldsymbol{z}_2)$.

2. *For any given $\boldsymbol{u}$, $p(y \mid \boldsymbol{u}, \boldsymbol{z})$ has a Lebesgue density with a monotone likelihood ratio.*

The variable $\boldsymbol{z}$ in the condition 1 of the above theorem is called an instrumental variable. When data analysts use this condition, it requires one to identify variables that affect outcomes but do not influence the response mechanism. This condition may be justified by domain knowledge or external related research. One can check the condition 2 for each model we use. We present examples that satisfy this condition in later chapters.

In practical analyses, identifying variables that satisfy the condition 1 can be challenging. Additionally, there is always the risk of relying on incorrect external studies or domain knowledge, leading to a false selection of variables. To overcome this challenge, Miao et al. (2016) provides the identification conditions without instrumental variables. Miao et al. (2016) consider the normal outcome model

$$Y \mid \boldsymbol{x} \sim N\{\mu(\boldsymbol{x}; \boldsymbol{\gamma}), \sigma^2\}, \tag{5}$$

and a generalized response mechanism

$$\pi(\boldsymbol{x}, y; \boldsymbol{\phi}) = F\{g(\boldsymbol{x}; \boldsymbol{\alpha}) + \beta y\}, \tag{6}$$

where the function $F : \mathbb{R} \to (0, 1]$ is a known strict monotone, the functions $\mu(\boldsymbol{x}; \boldsymbol{\gamma})$ and $g(\boldsymbol{x}; \boldsymbol{\alpha})$ have the known functional form and injective function with respect to $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$, respectively. Additionally, they use the following condition about the function F to identify the parameter

$$^\forall\delta > 0, \ \lim_{z \to -\infty} F(z)/e^{\delta z} = 0 \text{ or } \infty. \tag{7}$$

Condition (7) stipulates that the left tail decay rate of the response probability must not be exponential. While the Probit missing mechanism meets Condition (7), the Logistic missing mechanism fails to satisfy it because $\lim_{z \to -\infty}\{e^z/(1 + e^z)\}/e^z = 1$. The next theorem is the identifiability conditions given by Miao et al. (2016).

Lemma 2.3. *(Identification condition proposed by Miao et al. (2016)) Assume that the outcome model and the response model follow* (5) *and* (6)*, respectively. Then,*

1. *The absolute value of the parameter* $|\beta|$ *is identifiable;*
2. *The observed likelihood is identifiable if the sign of* β *is known;*
3. *The observed likelihood is identifiable if Condition* (7) *holds;*
4. *The observed likelihood is identifiable if the functions* $\mu(\boldsymbol{x};\boldsymbol{\gamma})$ *and* $g(\boldsymbol{x};\boldsymbol{\alpha})$ *are linearly uncorrelated, that is,* $a\mu(\cdot;\boldsymbol{\gamma}) + g(\cdot;\boldsymbol{\alpha}) \neq c$ *for nonzero vector* (a, b) *and for all* $\boldsymbol{\gamma}$ *and* $\boldsymbol{\alpha}$*.*

The condition 1 of Lemma 2.3 provides the identifiability of $\|\beta\|$ without any assumption. In the condition 2 of Lemma 2.3, knowing the sign of β means that the parameter β is identifiable from the condition 1 of Lemma 2.3. In other words, the identifiability of β is equivalent to that of the observed likelihood (1),

Note that Miao et al. (2016) gives the theorem in a generalized form of variance to $\sigma^2(\boldsymbol{x};\boldsymbol{\theta})$. Also, Miao et al. (2016) provides the identification conditions when the outcome is a normal mixture and the response model follows the robit model with an unknown degree of freedom.

2.4.2 Identifiability for generalized method of moments estimator

In the previous section, we provided identifiability conditions for cases where the observed likelihood is maximized. This section focuses on identifiability in estimation methods that do not rely on observed likelihood maximization.

Chang and Kott (2008) proposed a semiparametric estimator for $\boldsymbol{\phi}$ that is the solution of the following estimating equation

$$\sum_{i=1}^{n} \Gamma(\boldsymbol{x}_i, y_i, \delta_i; \boldsymbol{\phi}) = \sum_{i=1}^{n} \left\{1 - \frac{\delta_i}{\pi(\boldsymbol{x}, y; \boldsymbol{\phi})}\right\} g(\boldsymbol{x}_i; \boldsymbol{\phi}) = \mathbf{0}, \tag{8}$$

where $g = \{g_1(\boldsymbol{x}_i;\boldsymbol{\phi}), g_2(\boldsymbol{x}_i;\boldsymbol{\phi}), \cdots, g_k(\boldsymbol{x}_i;\boldsymbol{\phi})\}$ and k is dimension of $\boldsymbol{\phi}$. This estimator has consistency and asymptotic normality under appropriate regularity conditions, including the linearity independent of the function g with respect to $\boldsymbol{x}$. Certainly, identifiability is part of these regularity conditions. Morikawa and Kim (2021) derived the identification conditions for the parameter $\boldsymbol{\phi}_0$ defined below

$$E\{\Gamma(\boldsymbol{x}, y, \delta; \boldsymbol{\phi}) \mid \boldsymbol{x}\} = 0, \text{ a.s.}, \tag{9}$$

where the function Γ is defined in (8).

Lemma 2.4. *(Identification condition proposed by Morikawa and Kim (2021)) Let* $E_1(\cdot \mid \boldsymbol{x})$ *be expectation operator under the true conditional density function* $p(y \mid \boldsymbol{x}, \delta = 1)$ *and* $O(\boldsymbol{x}, y; \boldsymbol{\phi}) = 1/\pi(\boldsymbol{x}, y; \boldsymbol{\phi}) - 1$ *be the odds ratio of the response model. Then, the observed likelihood* (1) *is identifiable if the following conditions hold:*

1. $E_1\{O(\boldsymbol{x}, y; \boldsymbol{\phi}) \mid \boldsymbol{x}\}$ *is bounded almost surely;*
2. *The calibration function* g *in* (8) *satisfies* $P(\inf_{\boldsymbol{\phi}\in\Phi} |g(\boldsymbol{x};\boldsymbol{\phi})| > 0) > 0$*, and elements of* $g(\boldsymbol{x};\boldsymbol{\phi})$ *are linearly independent functions with respect to* $\boldsymbol{x}$ *for all* $\boldsymbol{\phi} \in \Phi$*.*

3. $E_1\left\{O(\boldsymbol{x}, y; \boldsymbol{\phi}) \mid x\right\} = E_1\left\{O(\boldsymbol{x}, y; \boldsymbol{\phi}') \mid x\right\}$ a.s. *implies* $\boldsymbol{\phi} = \boldsymbol{\phi}'$.

Condition 1 is a mild condition and also necessary to define the observed likelihood (1). Condition 2 requires to avoid g becomes identically zero. Condition 3 is essential to identify the parameter $\boldsymbol{\phi}$. Morikawa and Kim (2021) demonstrated the example that the response model is a logistic model. While condition 3 of Lemma 2.4 was established to identify the parameter (9), it is closely connected to the identifiability of the observed likelihood (1).

2.4.3 Identifiability for stability of inverse probability weighting methods

Li et al. (2023) consider a semiparametric location-scale model

$$y = \mu(\boldsymbol{x}; \boldsymbol{\gamma}) + \epsilon, \tag{10}$$

where the functional form of $\mu(\boldsymbol{x}; \boldsymbol{\gamma})$ is known, $\boldsymbol{\gamma}$ is a unknown parameter, and $E(\epsilon) = 0$. Also, they assume the logistic response model

$$P(R = 1 \mid \boldsymbol{x}, y) = \frac{\exp(\alpha_0 + \boldsymbol{x}^\top \boldsymbol{\alpha} + \beta y)}{1 + \exp(\alpha_0 + \boldsymbol{x}^\top \boldsymbol{\alpha} + \beta y)}, \tag{11}$$

where the parameter $\boldsymbol{\alpha}$ and β is unkown. Note that they do not assume the distribution of ϵ, thus, this is the semiparametric modeling. In this setup, they use an equation (4) from Li et al. (2023) to derive the following new logistic regression model:

$$P(R = 1 \mid \boldsymbol{x}) = \frac{1}{1 + \exp\left\{\alpha_0 + \boldsymbol{x}^\top \boldsymbol{\alpha} + c(\boldsymbol{x}; \beta, \boldsymbol{\gamma})\right\}}, \tag{12}$$

where $c(\boldsymbol{x}; \beta, \boldsymbol{\gamma}) = \log\left\{E\left(e^{\beta y} \mid \boldsymbol{x}, R = 1\right)\right\}$. Using this logistic regression model, they construct the likelihood

$$\mathrm{L}(\alpha_0, \boldsymbol{\alpha}, \beta, \hat{\boldsymbol{\gamma}}) = \prod_{R_i = 1} P(R_i = 1 \mid \boldsymbol{x}_i; \alpha_0, \boldsymbol{\alpha}, \beta, \hat{\boldsymbol{\gamma}}) \prod_{R_i = 0} P(R_i = 0 \mid \boldsymbol{x}_i; \alpha_0, \boldsymbol{\alpha}, \beta, \hat{\boldsymbol{\gamma}}), \tag{13}$$

and find the solution $\mathrm{argmax}_{\alpha_0, \boldsymbol{\alpha}, \beta} \mathrm{L}(\alpha_0, \boldsymbol{\alpha}, \beta, \hat{\boldsymbol{\gamma}})$. Note that to find the maximum likelihood estimator, they only use the data $\{(R_i, \boldsymbol{x}_i), i = 1, 2, \ldots, n\}$. Here, parameter identification refers to identifying the parameters in (13). In other words, from the equation (12), we should check that identifiability of $\alpha_0 + \boldsymbol{x}^\top \boldsymbol{\alpha} + c(\boldsymbol{x}; \beta, \boldsymbol{\gamma})$. Especially, we can obtain the specific form of the function $c(\boldsymbol{x}; \beta, \boldsymbol{\gamma})$ under the model (10) and (11) as follows

$$c(\boldsymbol{x}; \beta, \boldsymbol{\gamma}) = \beta \mu(\boldsymbol{x}; \boldsymbol{\gamma}) + \log M_1(\beta),$$

where $M_1(\beta) = E(e^{\beta \epsilon})$. Therefore, we reparameterize the model (12) to

$$P(R = 1 \mid \boldsymbol{x}) = \frac{1}{1 + \exp\left\{\xi + \boldsymbol{x}^\top \boldsymbol{\alpha} + \beta \mu(\boldsymbol{x}; \boldsymbol{\gamma})\right\}},$$

where $\xi = \alpha_0 + \log M_1(\beta)$, $(\xi, \boldsymbol{\alpha}, \beta)$ is a new parameter. Note that we should check the identifiability of $\xi + \boldsymbol{x}^\top \boldsymbol{\alpha} + \beta \mu(x; \boldsymbol{\gamma})$.

2.5 Proposed identification conditions: Discrete instrumental variables approach

Recently, widely applicable sufficient conditions have been proposed. Assume that a covariate $\boldsymbol{x}$ has two components, $\boldsymbol{x} = (\boldsymbol{u}^\top, \boldsymbol{z}^\top)^\top$, such that

(C1) $\boldsymbol{z} \perp\!\!\!\perp \delta \mid (\boldsymbol{u}, y)$ and $\boldsymbol{z} \not\perp\!\!\!\perp y \mid (\delta = 1, \boldsymbol{u})$.

The covariate $\boldsymbol{z}$ is called an instrument (D'Haultfœuille, 2010) or a shadow variable (Miao and Tchetgen Tchetgen, 2016). Miao et al. (2019) derived sufficient conditions for model identifiability by combining the instrument and the completeness condition:

(C2) For all square-integrable function $h(\boldsymbol{u}, y)$, $E[h(\boldsymbol{u}, y) \mid \delta = 1, \boldsymbol{u}, \boldsymbol{z}] = 0$ almost surely implies $h(\boldsymbol{u}, y) = 0$ almost surely.

Lemma 2.5 (Identification condition by Miao et al. (2019))**.** *Under the conditions (C1), (C2), the joint distribution $p(y, \boldsymbol{u}, \boldsymbol{z}, \delta)$ is identifiable.*

Although the completeness condition is useful and applicable for general models, a simple model with a categorical instrument does not hold the completeness condition.

Example 2.6 (Violating completeness with categorical instrument)**.** Suppose $y \mid (\delta = 1, u, z)$ follows the normal distribution $N(u + z, 1)$, the covariate u is one dimensional, and an instrument z is binary taking 0 or 1. This distribution does not satisfy the completeness condition because the conditional expectation $E[h(u, y) \mid \delta = 1, u, z] = 0$ when $h(u, y) = 1 + y - u - (y - u)^2$.

A vital implication of Example 2.6 is that instruments are no longer evidence of model identification when the instrument is categorical. Developing the identification condition for models with discrete instruments is important in applications (Ibrahim, Lipsitz and Horton, 2001). We separately discuss two cases: (i) both y and $\boldsymbol{z}$ are categorical; (ii) respondents' outcome model has the monotone-likelihood ratio property.

When all variables, y and $\boldsymbol{z}$, are categorical, the model can be fully nonparametric. Theorem 2.7 demonstrates that, under these conditions, the completeness and identifiability conditions are equivalent. See Appendix 2 in Riddles et al. (2016) for the estimation of such fully nonparametric models.

Theorem 2.7. *When both y and $\boldsymbol{z}$ are categorical, under condition (C1), the joint distribution $p(y, \boldsymbol{u}, \boldsymbol{z}, \delta)$ is identifiable if and only if condition (C2) holds.*

As evidenced in Lemma 2.5, condition (C2) is generally sufficient for model identifiability, but Theorem 2.7 also reveals that it is necessary when y and $\boldsymbol{z}$ are categorical.

Next, we consider the identification condition for the other case (ii). Let $\mathcal{S}_y$ be the support of the random variable y. We assume the following four conditions:

(C3) The response mechanism is

$$P(\delta = 1 \mid y, \boldsymbol{x}; \boldsymbol{\phi}) = P(\delta = 1 \mid y, \boldsymbol{u}; \boldsymbol{\phi}) = \Psi\{h(\boldsymbol{u}; \boldsymbol{\alpha}) + g(\boldsymbol{u}; \boldsymbol{\beta}) m(y)\}, \tag{14}$$

where $\boldsymbol{\phi} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta})^\top$, $m : \mathcal{S}_y \to \mathbb{R}$ and $\Psi : \mathbb{R} \to (0, 1]$ are known continuous strictly monotone functions, and $h(\boldsymbol{u}; \boldsymbol{\alpha})$ and $g(\boldsymbol{u}; \boldsymbol{\beta})$ are known injective functions of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively.

(C4) The density or mass function $p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma})$ is identifiable, and its support does not depend on $\boldsymbol{x}$.

(C5) For all $\boldsymbol{u} \in \mathcal{S}_{\boldsymbol{u}}$, there exist $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, such that $p(y \mid \boldsymbol{u}, \boldsymbol{z}_1, \delta = 1) \neq p(y \mid \boldsymbol{u}, \boldsymbol{z}_2, \delta = 1)$, and $p(y \mid \boldsymbol{u}, \boldsymbol{z}_1, \delta = 1)/p(y \mid \boldsymbol{u}, \boldsymbol{z}_2, \delta = 1)$ is monotone with respect to y.

(C6)

$$\int \frac{p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma})}{\Psi\{h(\boldsymbol{u}; \boldsymbol{\alpha}) + g(\boldsymbol{u}; \boldsymbol{\beta})m(y)\}} dy < \infty \quad \text{a.s.}$$

The condition (C3) means that the random variable $\boldsymbol{z}$ plays a role of an instrument. The condition (C4) is the identifiability of $p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma})$, which is testable from the observed data. The condition (C5) assumes a monotone-likelihood property on the outcome model, which was also used in Wang et al. (2014) for the complete data. The condition (C6) is necessary for (1) to be well-defined and essentially the same as condition 1 of Lemma 2.4. This condition is always true when the support of y is finite. However, it must be carefully verified when y is continuous. See Proposition 2.11 below for useful sufficient conditions when the respondents' outcome model is normal distribution.

Under conditions (C3)–(C6), we obtain the desired identification condition.

Theorem 2.8. *The parameter $(\boldsymbol{\phi}^\top, \boldsymbol{\gamma}^\top)^\top$ is identifiable if the conditions (C1) and (C3)–(C6) hold.*

We provide an example of outcome models satisfying the condition (C5).

Example 2.9 (Model satisfying (C5))**.** Let density functions in the exponential family be

$$p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma}) = \exp\left(\frac{y\theta - b(\theta)}{\tau} + c(y; \tau)\right),$$

where $\theta = \theta(\eta)$, $\eta = \sum_{l=1}^{L} \eta_l(\boldsymbol{x})\kappa_l$, $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_L)^\top$, and $\boldsymbol{\gamma} = (\tau, \boldsymbol{\kappa}^\top)^\top$. Then the density ratio becomes

$$\frac{p(y \mid \boldsymbol{u}, \boldsymbol{z}_1, \delta = 1)}{p(y \mid \boldsymbol{u}, \boldsymbol{z}_2, \delta = 1)} \propto \exp\left(\frac{\theta_1 - \theta_2}{\tau} y\right),$$

where $\boldsymbol{x}_i = (\boldsymbol{u}_i, \boldsymbol{z}_i)$ and $\theta_i = \theta\{\sum_{l=1}^{L} \eta_l(\boldsymbol{x}_i)\kappa_l\}$, $i = 1, 2$. Therefore, the density ratio is monotone.

Example 2.10 (Model satisfying (C6))**.** In application, it is often reasonable to assume a normal distribution on the respondents' outcome model. Focusing on the tail of the outcome model, we provide a sufficient condition to check (C6) for models with general response mechanisms.

Proposition 2.11. *Suppose that the observed distribution $p(y \mid \boldsymbol{x}, \delta = 1)$ is normal distribution $N(\mu(\boldsymbol{x}; \boldsymbol{\kappa}), \sigma^2)$, the response mechanism is (14) with $m(y) = y$ and $g(\boldsymbol{u}; \boldsymbol{\beta}) = \beta$, and the strictly monotone increasing function Ψ meets the following condition:*

$$^\exists s \in (0, 2) \text{ s.t. } \liminf_{z \to -\infty} \Psi(z) \exp(|z|^s) > 0. \tag{15}$$

Then, this model satisfies (C6).

The condition (15) is easy to check. For example, it holds for Logistic and Robit functions but not for the Probit function. According to Proposition 2.11, it is possible to estimate $\mu(x;\boldsymbol{\kappa})$ with observed data using splines and other nonparametric methods, which allows us to use very flexible models. Furthermore, we can also estimate the response mechanism using nonparametric methods because it does not impose any restrictions on the functional form of $h(\boldsymbol{u};\boldsymbol{\alpha})$.

2.6 Proposed identification conditions: Avoiding instrumental variables approach

In the previous section, we consider the identification conditions under the existence of instrumental variables and conduct various discussions. However, it is well known that the specification of such variables is challenging in real-world scenarios. Therefore, in this section, we explore identifiability without relying on the existence of instrumental variables. Note that this section is based on the findings of Beppu et al. (2024). First, we assume that the response mechanism follows a logistic model:

$$\operatorname{logit}\{P(\delta=1 \mid \boldsymbol{x}, y;\boldsymbol{\alpha},\beta)\} = h(\boldsymbol{x};\boldsymbol{\alpha}) + \beta y, \tag{16}$$

where $\operatorname{logit}(z) := \log(z/(1-z))$ for all $0<z<1$, $h(\boldsymbol{x};\boldsymbol{\alpha})$ is injective with respect to $\boldsymbol{\alpha}$, and is known up to a finite-dimensional parameter $\boldsymbol{\alpha}$. In nonignorable missing data analysis, several previous studies have employed this logistic model (Kim and Yu, 2011; Shao and Wang, 2016; Wang, Lu and Liu, 2021).

Suppose that the outcome variable distribution of the respondent, given the covariates $[y_i \mid \boldsymbol{x}_i, \delta_i = 1]$ belongs to the exponential family in the form

$$p(y_i \mid \boldsymbol{x}_i, \delta_i = 1;\boldsymbol{\gamma}) = \exp\left[\tau\{y_i\theta_i - b(\theta_i)\} + c(y_i;\tau)\right], \tag{17}$$

where $\theta_i = \theta(\eta_i)$, $\eta_i = \sum_{l=1}^{L} \eta_l(\boldsymbol{x}_i)\kappa_l$, $\boldsymbol{\kappa} = (\kappa_1,\ldots,\kappa_L)^\top$, and $\boldsymbol{\gamma} = (\tau, \boldsymbol{\kappa}^\top)^\top$. This class includes several distributions such as binomial, normal, gamma, and Poisson. Function θ is defined according to the purpose of the statistical analysis. In fact, this joint distribution can explicitly represent the distribution of the missing part $p(y \mid \boldsymbol{x}, \delta = 0)$. To discuss this in detail, when the outcome model of the respondents belongs to the exponential family in (17) and the response mechanism follows the logistic model in (16), the outcome model of the non-respondents also belongs to the same exponential family, but with a different parameterization:

$$\begin{aligned} p(y \mid \boldsymbol{x}, \delta = 0) &= p(y \mid \boldsymbol{x}, \delta = 1)\frac{\exp\{-h(\boldsymbol{x};\boldsymbol{\alpha}) - \beta y\}}{\int \exp\{-h(\boldsymbol{x};\boldsymbol{\alpha}) - \beta y\}\, p(y \mid \boldsymbol{x}, \delta = 1) dy} \\ &\propto \exp\left[\tau\left\{y\left(\theta - \beta\tau^{-1}\right) - b\left(\theta - \beta\tau^{-1}\right)\right\} + c(y;\tau)\right]. \end{aligned} \tag{18}$$

2.6.1 Single outcome model

In this section, we derive sufficient conditions representing model identifiability, considering the outcome models that belong to the exponential family (17), and extend the result to its mixture. The following theorem is a general result of model identifiability for the outcome model (17).

Theorem 2.12. *Suppose that the response mechanism is (16) and the distribution of* $[y \mid \boldsymbol{x}, \delta = 1]$ *is identifiable with a density that belongs to the exponential family (17). Then, this model is identifiable if and only if the following condition holds for all* $\boldsymbol{\alpha}, \boldsymbol{\alpha}', \beta, \beta', \boldsymbol{\gamma}$*:*

$$\varphi(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}) = \varphi(\boldsymbol{\alpha}', \beta', \boldsymbol{\gamma}) \Rightarrow \beta = \beta', \tag{19}$$

where

$$\varphi(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}) = h(\boldsymbol{x}; \boldsymbol{\alpha}) - \tau b \left\{ \theta \left(\sum_{l=1}^{L} \eta_l(\boldsymbol{x}) \kappa_l \right) - \frac{\beta}{\tau} \right\}.$$

A vital implication of Theorem 2.12 is that the identifiability of the model is equivalent to that of $\varphi(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma})$. Furthermore, the model identification of φ can be verified only with respect to β. Based on Theorem 2.12, we can conveniently check the identification conditions for almost all distributions belonging to the exponential family even if the covariates $\boldsymbol{x}$ contain both discrete and continuous variables.

Remark 1. We mention the relationship between Theorem 2.12 and Lemma 2.4. The difference lies in the estimation method. Specifically, Theorem 2.12 requires that the parameter maximizing the observed likelihood is unique, while Lemma 2.4 demands that the parameter uniquely provides the root of (9). However, the derived sufficient conditions are essentially the same. In other words, for models within the scope of both methods, the identification conditions are equivalent. A minor difference is that Theorem 2.12 is presented in a more specific form and the identification of β alone is sufficient in this context.

Remark 2. We consider the relationship between Theorem 2.12 and the identification condition of Li et al. (2023) presented in section 2.4.3. The primary difference is that Li et al. (2023) assumes a semiparametric model, whereas Theorem 2.12 is established under a fully parametric model. In scenarios where both are applicable, such as when the outcome is normally distributed, their identification conditions are equivalent. In practice, it would be reasonable to use Li et al. (2023) when the outcome variable y can take any value in the real numbers, and Theorem 2.12 when y takes discrete values or categorical cases.

When the covariates $\boldsymbol{x}$ contain only discrete variables, we can determine whether the number of unknown variables $(\boldsymbol{\alpha}^\top, \beta)^\top$ is less than or equal to the number of values taken by the covariates $\boldsymbol{x}$. Additionally, we provide Corollary 2.13, which specifically assumes that the outcome model follows a normal distribution because it requires careful attention, as detailed in Example 2.14.

Corollary 2.13. *Suppose that the response mechanism is (16),* $h(\boldsymbol{x}; \boldsymbol{\alpha})$ *is a polynomial, and the outcome model of respondent is* $N(\mu(\boldsymbol{x}; \boldsymbol{\gamma}), \sigma^2)$*, where the link function represents the identity* $\theta(\eta) = \eta$ *such that* $\mu(\boldsymbol{x}; \boldsymbol{\gamma}) = \sum_{l=1}^{L} \eta_l(\boldsymbol{x}) \kappa_l$*. Then, condition (19) holds if an index* $l = 1, \ldots, L$ *exists such that* $\eta_l(\boldsymbol{x})$ *is continuous and not represented by* $h(\boldsymbol{x}; \boldsymbol{\alpha})$ *for all* $\boldsymbol{\alpha}$*.*

For better understanding of Theorem 2.12 and Corollary 2.13, we introduce some examples that are commonly used in the generalised linear model (GLM).

Example 2.14. Considering the same setting as in Corollary 2.13, the function φ in Theorem 2.12 can be expressed as

$$\varphi(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}) = h(\boldsymbol{x}; \boldsymbol{\alpha}) + \beta \sum_{l=1}^{L} \eta_l(\boldsymbol{x})\kappa_l - \frac{\sigma^2 \beta^2}{2}.$$

Condition (19) holds if $\sum_{l=1}^{L} \eta_l(\boldsymbol{x})\kappa_l$ contains a term not included in $h(\boldsymbol{x}; \boldsymbol{\alpha})$. For instance, $\sum_{l=1}^{L} \eta_l(x)\kappa_l = \kappa_0 + \kappa_1 x + \kappa_2 x^2$ $(\kappa_2 \neq 0)$ and $h(x; \boldsymbol{\alpha}) = \alpha_0 + \alpha_1 x$ satisfy this condition if the covariate x is continuous and one dimension. If covariate x is binary, the identification condition does not hold because of the three unknown variables, $(\alpha_0, \alpha_1, \beta)$. In addition, the model is not identifiable for $\kappa_2 = 0$ even if covariate x is continuous. This is identical to the example in Morikawa and Kim (2021).

Example 2.15. Suppose $[y \mid \boldsymbol{x}, \delta = 1] \sim B(1, p(\boldsymbol{x}))$, which belongs to the exponential family, with $\tau = 1$, $\theta = \log p/(1-p)$, $b(\theta) = \log\{1 + \exp(\theta)\}$, $c(y; \tau) = 0$, $\theta(\eta) = \eta$. Accordingly, we check the identification of this model, and function φ in Theorem 2.12 can be expressed as

$$\varphi(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}) = h(\boldsymbol{x}; \boldsymbol{\alpha}) - \log\left\{1 + \exp\left(-\beta + \sum_{l=1}^{L} \eta_l(\boldsymbol{x})\kappa_l\right)\right\}.$$

For example, condition (19) holds if the polynomials $h(\boldsymbol{x}; \boldsymbol{\alpha})$ and $\eta_l(\boldsymbol{x})$ contain continuous variables.

For discrete nonmeasurement variables such as sex and area, the outcome models for each nonmeasurement variable should be assumed, as discussed in Section 5. For instance, if z denotes sex and x represents one-dimensional continuous covariate, we can model various mean structures: $\kappa_{01} + \kappa_{11} x$ for males and $\kappa_{02} + \kappa_{12} x$ for females rather than $\kappa_0 + \kappa_1 x + \kappa_2 z$. In these cases, we have sufficient conditions for model identifiability.

Example 2.16. Suppose that nonmeasurement categorical variables occur in D cases with z $(= 1, 2, \ldots, D)$ indicating one of the D cases, the response mechanism is (16) and $h(\boldsymbol{x}; \boldsymbol{\alpha}) = \sum_{d=1}^{D} h_d(\boldsymbol{x}; \boldsymbol{\alpha}_d)$, and the outcome density of the respondents can be expressed as

$$\prod_{d=1}^{D} \left[\frac{1}{\sqrt{2\pi\sigma_d^2}} \exp\left\{-\frac{(y - \mu_d(\boldsymbol{x}; \boldsymbol{\kappa}_d))^2}{2\sigma_d^2}\right\}\right]^{I(z=d)},$$

where $\boldsymbol{\kappa}_d$ and σ_d^2 represent the mean function and variance parameter, respectively, and $I(\cdot)$ denotes the indicator function. As it is normal distribution with mean $\sum_{d=1}^{D} I(z = d)\mu_d(\boldsymbol{x}; \boldsymbol{\kappa}_d)$ and variance $\sum_{d=1}^{D} I(z = d)\sigma_d^2$, according to Example 2.14, the function φ in Theorem 2.12 can be stated as

$$\varphi(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}) = \sum_{d=1}^{D} h_d(\boldsymbol{x}; \boldsymbol{\alpha}_d) + \beta I(z = d)\mu_d(\boldsymbol{x}; \boldsymbol{\kappa}_d) - \frac{\beta^2}{2} I(z = d)\sigma_d^2.$$

Similar to Corollary 2.13, this model is identifiable if an index $l = 1, \ldots, D$ exists, such that $\mu_l(\boldsymbol{x}; \boldsymbol{\kappa}_l)$ is continuous and not represented by $h_l(\boldsymbol{x}; \boldsymbol{\alpha}_l)$ for all $\boldsymbol{\alpha}_l$.

Example 2.17. Suppose that nonmeasurement categorical variables occur in D cases with z $(= 1, 2, \ldots, D)$ indicating one of the D cases, the response mechanism is (16) and $h(\boldsymbol{x}; \boldsymbol{\alpha}) = \sum_{d=1}^{D} h_d(\boldsymbol{x}; \boldsymbol{\alpha}_d)$, and $[y \mid \boldsymbol{x}, \delta = 1] \sim B(1, p(\boldsymbol{x}))$, where $\text{logit}\{p(\boldsymbol{x})\} = \sum_{d=1}^{D} I(z = d) \sum_{l=1}^{L_d} \eta_{ld}(\boldsymbol{x})\kappa_{ld}$. The function φ in Theorem 2.12 can be expressed as

$$\varphi(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}) = \sum_{d=1}^{D} h_d(\boldsymbol{x}; \boldsymbol{\alpha}_d) - \log \left\{ 1 + \exp\left(-\beta + \sum_{d=1}^{D} I(z = d) \sum_{l=1}^{L_d} \eta_{ld}(\boldsymbol{x})\kappa_{ld} \right) \right\}.$$

Similar to Example 2.16, the model can be identified more easily than in the case without categorical measurement variables.

2.6.2 Mixture outcome models

In this subsection, we derive sufficient conditions for model identifiability when the response mechanism is (16), and the outcome model of the respondents $[y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma}]$ is a mixture distribution of the exponential family (17)

$$\sum_{k=1}^{K} \pi_k \exp\left[\tau_k \left\{y\theta_k - b(\theta_k)\right\} + c(y; \tau_k)\right], \tag{20}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^\top$ represents the mixing proportion of the mixture models, i.e., $\sum_{i=k}^{K} \pi_k = 1$ and $\pi_k \geq 0$, $\theta_k = \theta(\eta_k)$ and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)^\top$ denote the model parameters, $\eta_k = \sum_{l=0}^{m(k)} \eta_{lk}(\boldsymbol{x})\kappa_{lk}$, $\boldsymbol{\kappa}_k = (\kappa_{0k}, \kappa_{1k}, \ldots, \kappa_{m(k)k})^\top$ and $\boldsymbol{\kappa} = (\boldsymbol{\kappa}_1^\top, \ldots, \boldsymbol{\kappa}_K^\top)^\top$ are link functions and their parameters, $\boldsymbol{\gamma} = (\kappa^\top, \tau^\top, \pi^\top, K)^\top$ is a vector of all of the parameters, and $m(k) + 1$ indicates a dimension of the vector $\boldsymbol{\kappa}_k$. The following theorem is the most general result representing the identifiability of the mixture model, and its results are consistent with those of Theorem 2.12 for $K = 1$.

Theorem 2.18. *Suppose that the response mechanism is (16) and the distribution of $[y \mid \boldsymbol{x}, \delta = 1]$ is (20) and identifiable. Then, this model is identifiable if and only if the following condition holds for all $\boldsymbol{\alpha}, \boldsymbol{\alpha}', \beta, \beta', \boldsymbol{\gamma}$:*

$$g(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}) = g(\boldsymbol{\alpha}', \beta', \boldsymbol{\gamma}) \Rightarrow \beta = \beta',$$

where

$$g(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}) = h(\boldsymbol{x}; \boldsymbol{\alpha}) - \log \left[\sum_{k=1}^{K} \pi_k \exp\left\{ -\tau_k b(\theta_k) + \tau_k b\left(\theta_k - \frac{\beta}{\tau_k} \right) \right\} \right].$$

Because the most popular and commonly used mixture model is a normal mixture, we discuss it in more detail the identification conditions for the case where $[y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma}]$ follows a normal mixture distribution:

$$[y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma}] \sim \sum_{k=1}^{K} \pi_k N(\mu_k(\boldsymbol{x}; \boldsymbol{\kappa}_k), \sigma_k^2), \tag{21}$$

where $\mu_k(\boldsymbol{x}; \boldsymbol{\kappa}_k)$ denotes a polynomial $\sum_{l=0}^{m(k)} \eta_{lk}(\boldsymbol{x})\kappa_{lk}$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_K^2)^\top$ represents a vector of variance, and $\boldsymbol{\gamma} = (\boldsymbol{\kappa}^\top, {\boldsymbol{\sigma}^2}^\top, \boldsymbol{\pi}^\top, K)^\top$ denotes a vector of all of the parameters. In this case, we obtain Corollary 2.19 by applying Theorem 2.18:

Corollary 2.19. *Suppose that the response mechanism is (16) and the distribution of* $[y \mid \boldsymbol{x}, \delta = 1]$ *is (21) and identifiable. Then, this model is identifiable if and only if the following condition holds for all* $\boldsymbol{\alpha}, \boldsymbol{\alpha}', \beta, \beta', \boldsymbol{\gamma}$*:*

$$g(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}) = g(\boldsymbol{\alpha}', \beta', \boldsymbol{\gamma}) \Rightarrow \beta = \beta',$$

where

$$g(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}) = h(\boldsymbol{x}; \boldsymbol{\alpha}) - \log\left\{\sum_{k=1}^{K} \pi_k \exp\left(-\beta \sum_{l=0}^{m(k)} \eta_{lk}(\boldsymbol{x})\kappa_{lk} + \frac{\beta^2\sigma_k^2}{2}\right)\right\}.$$

Hereafter, we consider a practically useful setup, where $h(x; \boldsymbol{\alpha}) = \sum_{j=0}^{J-1} \alpha_j h_j(\boldsymbol{x})$ and $\mu_k(x; \boldsymbol{\kappa}_k) = \sum_{l=0}^{m(k)} \eta_{lk}(\boldsymbol{x})\kappa_{lk}$, where J is the dimension of $\boldsymbol{\alpha}$, the basis functions of $h_j(\boldsymbol{x})$ and $\eta_{lk}(\boldsymbol{x})$ have the form of polynomial function $\prod_{i=1}^{\dim(\boldsymbol{x})} x_i^{s_i}$, $\dim(\boldsymbol{x})$ is a dimensional $\boldsymbol{x}$, and s_i represents any nonnegative integer. We define two classes of basis functions:

$$\begin{aligned}
\mathcal{H} &:= \left\{h_j(\boldsymbol{x}) \,\middle|\, j = 0, 1, \ldots, J-1\right\} \cup \{1\}, \\
\mathcal{M} &:= \left\{\eta_{lk}(\boldsymbol{x}), \,\middle|\, l = 0, 1, \ldots, m(k),\ k = 1, \ldots, K\right\} \setminus \mathcal{H},
\end{aligned}$$

and decompose $\mu_k(\boldsymbol{x}; \boldsymbol{\kappa}_k)$ into $\mu_k(\boldsymbol{x}; \boldsymbol{\kappa}_k) = \mu_k^{\mathcal{H}}(\boldsymbol{x}; \boldsymbol{\kappa}_k) + \mu_k^{\mathcal{M}}(\boldsymbol{x}; \boldsymbol{\kappa}_k)$, where each $\mu_k^{\mathcal{H}}$ and $\mu_k^{\mathcal{M}}$ are constant multiple of the elements of $\mathcal{H}$ and $\mathcal{M}$, respectively. For example, in the case of $h(x; \boldsymbol{\alpha}) = 2 + 4x$, $\mu_1(x; \boldsymbol{\kappa}_1) = 3x^2$, and $\mu_2(x; \boldsymbol{\kappa}_2) = x + 4x^3$, the definitions of the notation imply $\mathcal{H} = \{1, x\}$, $\mathcal{M} = \{x^2, x^3\}$, $\mu_1^{\mathcal{H}} = 0$, $\mu_1^{\mathcal{M}} = 3x^2$, $\mu_2^{\mathcal{H}} = x$, and $\mu_2^{\mathcal{M}} = 4x^3$.

When the distribution of $\boldsymbol{x}$ is discrete, comparing the number of unknown variables and the taken values of $\boldsymbol{x}$ is sufficient. Therefore, we consider only the continuous case:

(C7) The distribution of $\boldsymbol{x}$ is continuous.

The next theorem provides more rigorous conditions and identifiability results given the above setting.

Theorem 2.20. *Suppose that the response mechanism is (16) and the distribution of* $[y \mid \boldsymbol{x}, \delta = 1]$ *is identifiable and has a normal mixture density in (21). Furthermore, we define three additional conditions:*

(C8) $\mathcal{M} \neq \emptyset$*;*

(C9) The sign of β *is known;*

(C10) $\left\{\mu_i^{\mathcal{M}}(\boldsymbol{x}; \boldsymbol{\kappa}_i);\ i = 1, \ldots, K\right\} \neq \left\{-\mu_i^{\mathcal{M}}(\boldsymbol{x}; \boldsymbol{\kappa}_i);\ i = 1, \ldots, K\right\}$.

Then, this model is identifiable if (C7)–(C8) and one of (C9)–(C10) hold.

In application, confirming (C8) may be sufficient because violation of (C10) is rare in practical applications.

The following example shows an unidentifiable model that satisfies the condition (C8), but does not satisfy (C9)–(C10).

Example 2.21. Suppose the outcome model for respondents is $\pi_1 N(x^2, \sigma_1^2) + \pi_2 N(-x^2, \sigma_2^2)$, where x is one-dimensional covariate. Consider the following two response models:

$$\begin{aligned} \text{Model 1}: \ & \text{logit}\{P(\delta = 1 \mid x, y)\} = x + y; \\ \text{Model 2}: \ & \text{logit}\{P(\delta = 1 \mid x, y)\} = x - y. \end{aligned}$$

This model satisfies condition (C8) because $\mathcal{M} = \{x^2\}$ and $\mathcal{H} = \{1, x\}$, but does not satisfy (C9)–(C10) because we do not know the sign of β and $\{\mu_i^{\mathcal{M}}(\boldsymbol{x}; \boldsymbol{\kappa}_i);\ i = 1, \ldots, K\} = \{x^2, -x^2\}$. The sufficient condition in Corollary 2.19 is not satisfied if

$$\frac{1}{2}\sigma_1^2 + \log \pi_1 = \frac{1}{2}\sigma_2^2 + \log \pi_2$$

holds; thus, this model is unidentifiable.

The following example shows an unidentifiable model that does not satisfy (C8).

Example 2.22. Suppose that the outcome model of the respondents is $\pi_1 N(x, \sigma_1^2) + \pi_2 N(2x, \sigma_2^2)$, where x is one-dimensional covariate. Consider two response models:

$$\begin{aligned} \text{Model 1}: \ & \text{logit}\{P(\delta = 1 \mid x, y)\} = x + y; \\ \text{Model 2}: \ & \text{logit}\{P(\delta = 1 \mid x, y)\} = 4x - y. \end{aligned}$$

This model does not satisfy the condition (C8) because $\mathcal{M} = \emptyset$ and $\mathcal{H} = \{1, x\}$. The sufficient condition in Corollary 2.19 is not satisfied if

$$\frac{1}{2}\sigma_1^2 + \log \pi_1 = \frac{1}{2}\sigma_2^2 + \log \pi_2$$

holds; thus, this model is unidentifiable.

Although Example 2.22 and Theorem 2.20 indicate the importance of condition (C8), eliminating this condition enables a more flexible model. Thus, we derive sufficient conditions for the identifiability of a mixture of simple linear regression models:

$$[y \mid x, \delta = 1; \boldsymbol{\gamma}] \sim \sum_{k=1}^{K} \pi_k N(\kappa_{0k} + \kappa_{1k} x, \sigma_k^2), \tag{22}$$

and $\mathcal{H} = \{1, x\}$, which do not satisfy (C8). Because the sufficient conditions for model identifiability differ for $K \geq 3$ and $K = 2$, we derive the conditions separately.

Theorem 2.23. *Suppose that the response mechanism is (16) with $h(x; \boldsymbol{\alpha}) = \alpha_0 + \alpha_1 x$ and the distribution of $[y \mid x, \delta = 1]$ is identifiable and has a normal mixture density in (22) with the number of mixture components $K \geq 3$. We further assume that $\mathcal{H} = \{1, x\}$, $\tilde{\boldsymbol{\kappa}} = (\kappa_{11}, \kappa_{12}, \ldots, \kappa_{1K})^\top$ is a vector of first-order coefficients, and $\kappa_{1i} \neq \kappa_{1j}$ $(i \neq j)$. The model is identifiable if at least one of the following conditions is satisfied:*

1. *Sign of β is known;*

2. *For all $K \times K$ permutation matrices P and for all $r \in \mathbb{R}$,*

$$(P + I)\, \tilde{\boldsymbol{\kappa}} \neq r \mathbf{1}_K,$$

where $\mathbf{1}_n$ denotes an $n \times 1$ vector of ones.

To clarify the second condition, we consider $K = 3$ and assume that $\kappa_{11} > \kappa_{12} > \kappa_{13}$ without loss of generality. If P is defined as

$$P = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix},$$

we obtain $(P+I)\tilde{\boldsymbol{\kappa}} = (\kappa_{11}+\kappa_{13}, 2\kappa_{12}, \kappa_{11}+\kappa_{13})^\top$. Then, the second condition is satisfied, unless $2\kappa_{12} = \kappa_{11} + \kappa_{13}$. More importantly, it does not generally hold and can be tested using the observed data.

However, for $K = 2$, the second condition in Theorem 2.23 does not hold for any model in $[y \mid \boldsymbol{x}, \delta = 1]$, which can be demonstrated by assuming $\kappa_{11} > \kappa_{12}$ without loss of generality. Using a similar argument, we derive $(P+I)\tilde{\boldsymbol{\kappa}} = (\kappa_{11}+\kappa_{12}, \kappa_{11}+\kappa_{12})^\top = r\mathbf{1}_2$, where

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \; r = \kappa_{11} + \kappa_{12}.$$

Thus, the second condition does not hold for any model $[y \mid \boldsymbol{x}, \delta = 1]$. Therefore, a more careful investigation is required for $K = 2$.

Theorem 2.24. *Suppose that the response mechanism is (16) and the distribution of $[y \mid x, \delta = 1]$ is identifiable and has a normal mixture density in (22) with the number of mixture components $K = 2$. We further assume that $\mathcal{H} = \{1, x\}$, $\mu_1(x, \boldsymbol{\kappa}_1) = \kappa_{01} + \kappa_{11}x$ $(\kappa_{11} \neq 0)$, $\mu_2(x, \boldsymbol{\kappa}_2) = \kappa_{02} + \kappa_{12}x$ $(\kappa_{12} \neq 0)$, and $h(\boldsymbol{x}; \boldsymbol{\alpha}) = \alpha_0 + \alpha_1 x$ $(\alpha_1 \neq 0)$. The model is then identifiable if the following conditions hold:*

1 $\kappa_{11} \neq \kappa_{12}$;

2 $\sigma_1 = \sigma_2 \Rightarrow \pi_1 \neq \pi_2$;

3 $\sigma_1 \neq \sigma_2 \Rightarrow \; (\log \pi_2 - \log \pi_1)\,(\sigma_1^2 - \sigma_2^2)^{-1} \leq 0.$

Overall, the conditions required in Theorems 2.24 are more difficult to satisfy than those in Theorems 2.23. Although these conditions can be verified using the observed data, they are redundant.

3 Nonignorable model in survey sampling

In this section, the discussion transitions to the survey sampling, focusing on informative sampling, which is closely related to nonignorable mechanisms. Here, we extend the concept of double robustness from previous research to multiple robustness by leveraging empirical weights. Additionally, as a direction for future work, we introduce the issue of nonignorable participation in non-probability survey samples.

3.1 Efficient multiple robust estimation under informative sampling

3.1.1 Notation

Suppose that there exists a superpopulation of random variables (X, Y, Z, W), where Y denotes an outcome, X and Z are explanatory variables, and W represents a sampling weight. Specifically, the sampling weight is the inverse of the inclusion probability. We draw identically and independently distributed N copies $\{X_i, Y_i, Z_i, W_i\}_{i=1}^N$ from the distribution. Our aim is to estimate a parameter θ that characterizes the relationship between X and Y. The target parameter θ^* is uniquely determined by the solution to the equation $E\{U_\theta(X, Y)\} = 0$. For example, if our focus is on $E(Y)$, then $U_\theta(y) = y - \theta$, and if our interest lies in the regression parameter θ within $\mu(x; \theta) = E(Y \mid x; \theta)$, then $U_\theta(x, y) = A(x)\{y - \mu(x; \theta)\}$, where $A(x)$ is a function of x with the same dimensionality as θ.

In survey sampling, we extract a finite population of size n $(< N)$ based on the inclusion probability $1/W$. We Define δ as a sampling indicator that is set to one if a unit is sampled and set to zero otherwise. By the definition of the inclusion probability, $W = 1/P(\delta = 1 \mid X, Y, Z, W)$ holds. Consider a scenario in which the outcome variable Y is subject to missingness among the sampled units, resulting in only m out of n units having fully observed data. Let R be a response indicator for Y, which is equal to one when Y is observed and equal to zero otherwise.

We outline our setup across three distinct settings in Figure 1. In Setting 1, information regarding X is available for all units. In Setting 2, information regarding X is limited to sampled units. In Setting 3, instead of using data from X for unsampled units, we leverage additional data sources that provide information such as the mean and variance of X. Note that it is also possible to integrate external summary statistics into Setting 1 similar to Setting 3. For notational simplicity, without loss of generality, we rearrange the order of the units as follows: the initial m units are fully observed, the subsequent $n - m$ units are sampled with Y units missing, and the remaining $N - n$ units are unsampled.

We denote the independence of two random variables X and Y as $X \perp\!\!\!\perp Y$, and postulate the following two assumptions regarding the distribution of the superpopulation of (X, Y, Z, W, δ, R):

Condition 1 (Informative sampling)**.** *$W \not\perp\!\!\!\perp (Y, Z) \mid X$ in Setting 1; $W \not\perp\!\!\!\perp (X, Y, Z)$ in Settings 2 and 3;*

Condition 2 (Sample missing at random)**.** $R \perp\!\!\!\perp Y \mid (X, Z, W, \delta = 1)$.

When the negation of Condition 1 holds in Setting 1, we have

$$P(\delta = 1 \mid x, y, z) = E(W^{-1} \mid x, y, z) = E(W^{-1} \mid x) = P(\delta = 1 \mid x).$$

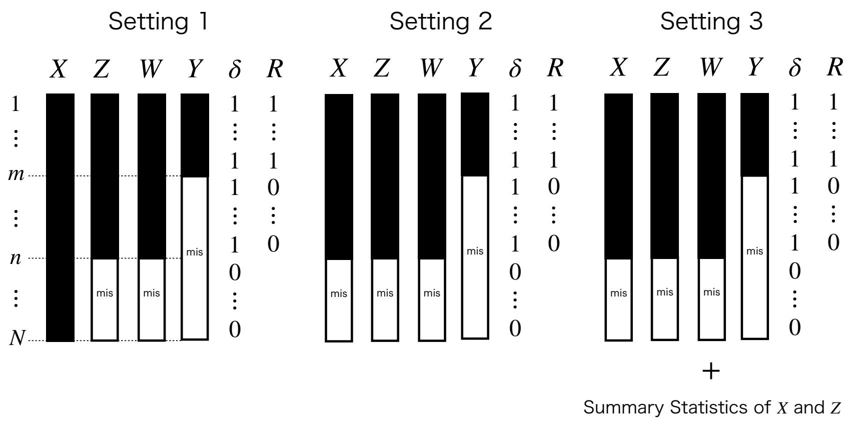

Figure 1: Three settings considered in this study. The data highlighted in black represents observed data, and the entries labeled as "mis" indicate unsampled or nonresponse.

This sampling mechanism is referred to as non-informative sampling in the literature of survey sampling (Pfeffermann, 1993) and as missing at random in missing data analysis (Rubin, 1976). Under this sampling mechanism, we have $f(y \mid x, \delta = 0) = f(y \mid x)$. Therefore, the sampled outcome distribution is the same as that of the population. In contrast, Condition 1 allows the two distributions to differ, and ignoring the sampling mechanism may result in biased results. This sampling mechanism is called informative or missing not at random. Condition 2 constrains the response mechanism of the subsequent missingness to be missing at random (Little, 2003; Pfeffermann, 1993), which differs from the population missing at randomness $R \perp\!\!\!\perp Y \mid X$ in Setting 1 defined in Berg, Kim and Skinner (2016). Condition 2 permits the response indicator to be dependent on sampling weights, which could be perceived as unconventional. Nevertheless, including sampling weights can enhance efficiency and mitigate bias resulting from design variables that data analysts cannot access, as these weights are related to the unsampled items and indirectly convey the information to the response indicator. As illustrated in Figure 1, our setup is essentially the same as that of two-step monotone missing data (e.g. Särndal, 1992), but the response mechanism combines missing not at random and missing at random, which makes statistical inference problematic (Kim and Shao, 2013; Little and Rubin, 2019).

3.1.2 Estimators from previous studies under informative sampling with non-response

Under Condition 2, if we know the response mechanism $\pi(x, z, w) = P(R = 1 \mid x, z, w, \delta = 1)$, then θ can be estimated by using the inverse probability weighted estimating equation (Binder, 1983):

$$\sum_{i=1}^{N} \frac{\delta_i W_i R_i}{\pi(X_i, Z_i, W_i)} U_\theta(X_i, Y_i) = \sum_{i=1}^{m} \frac{W_i}{\pi(X_i, Z_i, W_i)} U_\theta(X_i, Y_i) = 0. \tag{23}$$

However, the response mechanism is generally unknown and must be modeled and estimated. The parameters of the response model can be estimated using any method, such as maximum-likelihood estimation because X is observed for all units of $\delta = 1$. Because misspecification of the response model leads to biased results, we construct a double-robust estimator:

$$\sum_{i=1}^{N} \delta_i W_i \left[\frac{R_i}{\hat{\pi}(X_i, Z_i, W_i)} U_\theta(X_i, Y_i) + \left\{ 1 - \frac{R_i}{\hat{\pi}(X_i, Z_i, W_i)} \right\} \hat{g}_\theta(X_i, Z_i, W_i) \right] = 0, \tag{24}$$

where $g_\theta(x, z, w) = E\{U_\theta(x, Y) \mid x, z, w\}$ and "hat" denote estimated parametric models based on observed data (e.g., $\hat{\pi}(x, z, w)$ and $\hat{g}_\theta(x, z, w)$ are estimated functions using $\pi(x, z, w)$ and $g_\theta(x, z, w)$, where the detailed procedure is discussed in end of this section). Kim and Haziza (2014) considered a similar estimator without including weights in the response and the regression models. Although nonparametric working models are also usable due to orthogonality to the nuisance parameter (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins, 2018), we focus on parametric working models in this paper. The solution $\hat{\theta}_{\mathrm{KH}}$ to the estimating equation (24) has double robustness: if either $\pi(x, z, w)$ or $g_\theta(x, z, w)$ is correctly specified, then it has consistency. However, this estimator is not necessarily efficient, even if both models are correct, unlike the ordinal double-robust estimator (Robins, Rotnitzky and Zhao, 1994*b*), because it does not leverage the information in the data of $\delta_i = 0$ (i.e., X_i $(i = n+1, \ldots, N)$) in Setting 1 or the information of N in Settings 2 and 3.

Hereafter, we explain the semiparametric efficiency bound and the efficient score of the parameter θ in Settings 1 and 2 in Figure 1 derived by Morikawa et al. (2023).

Lemma 3.1. *The efficient score function in Setting 1 is*

$$S_{\mathrm{eff},\theta}(\delta, R, X, Y, Z, W) = \delta W D_\theta(R, X, Y, Z, W) + (1 - \delta W) C_\theta(X), \tag{25}$$

where

$$C_\theta(x) = \frac{E\{(W-1)U_\theta(x, Y) \mid x\}}{E(W - 1 \mid x)},$$

$$D_\theta(r, x, y, z, w) = r\frac{U_\theta(x, y)}{\pi(x, z, w)} + \left\{ 1 - \frac{r}{\pi(x, z, w)} \right\} g_\theta(x, z, w).$$

The efficient score function in Setting 2 is (25) with the same D_θ as above but different $C_\theta(x) = C_\theta = E\{(W-1)U_\theta(X, Y)\}/E(W-1)$. The efficient influence function is $\varphi_{\mathrm{eff},\theta} = B_\theta S_{\mathrm{eff},\theta}$, and the semiparametric efficiency bound for θ is $\{E(\varphi_{\mathrm{eff},\theta^*}^{\otimes 2})\}^{-1}$, where $B_\theta = E\{\partial U_\theta(X, Y)/\partial \theta^\top\}^{-1}$ and $B^{\otimes 2} = BB^\top$ for any matrix B.

To construct adaptive estimators, we require some working models for $\pi(x, z, w; \alpha)$, $g_\theta(x, z, w; \beta)$, and $C_\theta(X; \gamma)$, where α, β, and γ are unknown finite-dimensional parameters. Each model is naturally estimated based on the ignorability (Condition 2): $\hat{\alpha}$ is the maximum likelihood estimator,

$$\hat{\alpha} = \arg\max_{\alpha} \prod_{i=1}^{n} \pi(X_i, Z_i, W_i; \alpha)^{R_i} \{1 - \pi(X_i, Z_i, W_i; \alpha)\}^{1-R_i},$$

and $\hat{\beta}$ and $\hat{\gamma}$ are the weighted least-squares estimators

$$\sum_{i=1}^{m} B(X_i, Z_i, W_i)\{Y_i - g_\theta(X_i, Z_i, W_i; \beta)\}^2 = 0,$$

$$\sum_{i=1}^{m} W_i(W_i - 1)\{Y_i - C_\theta(X_i; \gamma)\}^2 = 0,$$

where $B(x, z, w)$ is any function of x, z, w that has the same dimensionality as β. In Setting 1, we define a method of moments estimator $\hat{\theta}_{\mathrm{MM1}}$ as the solution to

$$S^{[1]}_{\mathrm{eff},\theta}(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) = n^{-1} \sum_{i=1}^{N} \left\{ \delta_i W_i \hat{D}_\theta(R_i, X_i, Y_i, Z_i, W_i) + (1 - \delta_i W_i) \hat{C}_\theta(X_i) \right\} = 0, \tag{26}$$

whereas in Setting 2, we define $\hat{\theta}_{\mathrm{MM2}}$ by the solution to an estimating equation obtained by replacing $\hat{C}_\theta(x)$ with $\hat{C}_\theta$.

3.1.3 Derivation of the efficiency bound under Setting 3

In this section, we derive the efficiency bound in Setting 3. Note that sections 3.1.3, 3.1.4, and 3.1.5 are part of the contributions of my doctoral thesis and are included in Morikawa et al. (2023). In Setting 3, to make summary statistics from additional data sources available, we must extend the class of estimators. Let $P_0 \in \mathcal{P}_0$ and $P_1 \in \mathcal{P}_1$ be the internal and external distributions, respectively, where $\mathcal{P}_0$ and $\mathcal{P}_1$ are collections of probability distributions. We define the summary statistics $\tau^* = \tau(P_1)$ in the external data source. Additionally, we denote the estimator with an external data source of sample size N_1 as $\tilde{\tau} = \tilde{\tau}(P_1)$, which emphasizes that the parameter is estimated from the external distribution P_1. Integrating multiple data sources is also possible, but we confine our analysis to a single data source for the sake of simplicity. Readers can consult the discussion in Hu et al. (2022) for additional details on the extension to multiple data sources.

Following Hu et al. (2022), we adopt three assumptions regarding the summary statistics from the external data source.

Condition 3. *The summary statistics $\tau = \tau(P_1)$ represent a parameter of the fully observed covariates X and Z;*

Condition 4. *The estimated summary statistics $\tilde{\tau}$ represent a regular and asymptotically linear estimator of $\tau(P_1)$, and $N_1^{1/2}\{\tilde{\tau} - \tau(P_1)\}$ converges weakly to the normal distribution with mean zero and variance Σ_1, where Σ_1 is the asymptotic variance of $\tilde{\tau}$ and a consistent estimator $\tilde{\Sigma}_1$ for Σ_1 is available;*

Condition 5. *The sample size from the external data source $N_1 = N_{1,N}$ satisfies*

$$N_{1,N}/N \to \rho \in (0, \infty) \text{ as } N \to \infty;$$

Condition 6. $\tau(P_0) = \tau(P_1)$.

Condition 3 confines the summary statistics to the sampled data without missingness, such as sample means of X and Z, and regression coefficients Z on X, and excludes statistics related to outcome variables. Condition 4 is a regular condition for estimators. One may feel that Condition 5 is strange because internal and external data are often independent, but this assumption requires the sample size to go to infinity according to the ratio ρ. This condition is necessary to investigate the large-sample property and reflect the difference between the sample sizes N_1 and N. Condition 6 requires the consistency of the target parameter between two data sources. Specifically, the populations can differ, but the target parameters must be the same.

Let $I_i = (X_i, Y_i, Z_i, W_i, \delta_i, R_i)$ $(i = 1, \ldots, N)$ and E_i $(i = 1, \ldots, N_1)$ be random vectors in the internal and the external dataset, respectively. Then, our estimator for θ can be represented as $\hat{\theta}_N = \hat{\theta}_N(I_1, \ldots, I_N, \tilde{\tau})$ because it depends on both the internal data $\{I_i\}_{i=1}^N$ and summary statistics $\tilde{\tau}$ from the external data. We assume that our estimator in Setting 3 is in the class of data-fused regular and asymptotically linear estimators (Hu et al., 2022).

Definition 3.2 (Data-Fused Regular and Asymptotically Linear Estimator)**.** An estimator $\hat{\theta}_N = \hat{\theta}_N(I_1, \ldots, I_N, \tilde{\tau})$ is said to be data-fused regular and asymptotically linear if the following two conditions hold:

(i) (Regular). Let ξ be a finite-dimensional parameter in any parametric sub-model $P_0(I; \xi) \times P_1(E; \xi) \in \mathcal{P}_0 \times \mathcal{P}_1$ and ξ^* be the true value. Then, the variable

$$N^{1/2}\{\hat{\theta}_N(I_1^{(N)}, \ldots, I_N^{(N)}, \tilde{\tau}^{(N_1)}) - \theta(P_0(I; \xi_n))\}$$

has a limiting distribution that does not depend on the local data generation process, in which the data $\{I_1^{(N)}, \ldots, I_N^{(N)}\}$ and $\{E_1^{(N_1)}, \ldots, E_{N_1}^{(N_1)}\}$ are i.i.d. samples from $P_0(I; \xi_n)$ and $P_1(E; \xi_n)$, $\tilde{\tau}^{(N_1)}$ denotes the summary statistics estimated from $\{E_1^{(N_1)}, \ldots, E_{N_1}^{(N_1)}\}$, $N_1/N \to \rho$, and $N^{1/2}(\xi_n - \xi^*)$ converges to a constant.

(ii) (Asymptotically linear). The estimator $\hat{\theta}_N$ has the form $\hat{\theta}_N = \theta^* + N^{-1}\sum_{i=1}^N \psi_0(I_i) + \psi_1(\tilde{\tau}) + o_p(N^{-1/2})$, where $E\{\psi_0(I)\} = 0$, $E\{\psi_0(I)^{\otimes 2}\}$ is finite and non-singular, $\psi_1(\tilde{\tau})$ is continuously differentiable, and $\psi_1(\tau(P_1)) = 0$.

Hájek (1970) and Inagaki (1970) derived the convolution theorem independently, and it has been extended in various settings since then. Hu et al. (2022) extended the convolution theorem for the ordinary class of regular and asymptotically linear estimators to data-fused regular and asymptotically linear estimators and derived the efficiency bound of the new class.

Lemma 3.3. *Suppose $\hat{\theta}_N$ is a data-fused regular and asymptotically linear estimator and Conditions 3-6 hold. Then, we have*

$$N^{1/2}\begin{pmatrix} \hat{\theta}_N - \theta^* - N^{-1}\sum_{i=1}^N(\phi_{\mathrm{eff},i} - M\eta_{\mathrm{eff},i}) - M(\tilde{\tau} - \tau) \\ N^{-1}\sum_{i=1}^N(\phi_{\mathrm{eff},i} - M\eta_{\mathrm{eff},i}) + M(\tilde{\tau} - \tau) \end{pmatrix} \to \begin{pmatrix} \Delta_0 \\ \Delta_1 \end{pmatrix}$$

in distribution, where $M = E(\phi_{\mathrm{eff}}\eta_{\mathrm{eff}}^\top)\{\Sigma_1/\rho + E(\eta_{\mathrm{eff}}^{\otimes 2})\}^{-1}$, Δ_0 and Δ_1 are independent random variables, and $\phi_{\mathrm{eff},i}$ and $\eta_{\mathrm{eff},i}$ $(i = 1, \ldots, N)$ are efficient influence functions for θ and τ based on the internal data.

The convolution theorem yields the semiparametric efficiency bound in Setting 3 as $E(\phi_{\mathrm{eff}}^{\otimes 2}) - E(\phi_{\mathrm{eff}}\eta_{\mathrm{eff}}^\top)\{\Sigma_1/\rho + E(\eta_{\mathrm{eff}}^{\otimes 2})\}^{-1}E(\phi_{\mathrm{eff}}\eta_{\mathrm{eff}}^\top)^\top$ in the class of data-fused regular and asymptotically linear estimators. The first term is the efficiency bound when using only internal data. This implies that incorporating external summary statistics into estimation always results in a more efficient estimator.

3.1.4 Optimal estimator using empirical likelihood in Settings 1 and 2

We consider empirical likelihood based estimators with the same asymptotic variances detailed in Morikawa et al. (2023) for Settings 1 and 2. There are two primary motivations for considering the empirical likelihood methods. First, using the empirical likelihood method facilitates the easy derivation of multiple-robust estimators. Second, an elementary extension of the optimal empirical likelihood estimator from Setting 2 can provide the optimal estimator for Setting 3. Although it may be feasible to construct an estimator by the method of moments and extend it to Setting 3, we lean toward empirical likelihood estimators for their theoretically appealing properties; see Remark 4 for additional details.

One natural empirical likelihood estimator based on the efficient score (25) is defined as

$$\hat{\theta}_{\mathrm{EL,Q}} = \arg\max_{\theta} \arg\max_{p_1,\ldots,p_N} \sum_{i=1}^{N} \log p_i,$$

subject to

$$\sum_{i=1}^{N} p_i = 1,\ \sum_{i=1}^{N} p_i \delta_i W_i D_\theta(R_i, X_i, Y_i, Z_i, W_i) = 0,\ \sum_{i=1}^{N} p_i (1-\delta_i W_i) C_\theta(X_i) = 0.$$

By directly applying the theory in Qin, Zhang and Leung (2009), we can easily demonstrate that this maximum empirical likelihood estimator has desirable asymptotic properties. However, we do not adopt $\hat{\theta}_{\mathrm{EL,Q}}$ as our estimator because the function $D_\theta(r, x, y, z, w)$ includes the two working models $\pi(x, z, w)$ and $g(x, z, w)$, making it intractable to construct a multiple-robust estimator. Let our candidate models for $\pi(x, z, w)$ and $g_\theta(x, z, w)$ be $\pi^{[j]}(x, z, w)$ $(j = 1, \ldots . J)$ and $g_\theta^{[k]}(x, z, w)$ $(k = 1, \ldots, K)$, respectively. Then, $\hat{\theta}_{\mathrm{EL,Q}}$ requires JK constraints to obtain multiple robustness. Hereafter, in Settings 1 to 3, we propose two-step maximum empirical likelihood estimators to realize multiple robustness under only $J + K$ constraints. We also present the use of $C_\theta^{[l]}$ $(l = 1, \ldots, L)$ working models for $C_\theta(x)$ to achieve the efficiency bound.

The first-step empirical likelihood weights are common in all settings. We define the maximum empirical weights in the first step as

$$\hat{\theta}_1 = \arg\max_{\theta} \arg\max_{p_1^{(1)},\ldots,p_m^{(1)}} \sum_{i=1}^{m} \log p_i^{(1)},$$

subject to

$$\sum_{i=1}^{m} p_i^{(1)} = 1,\ \sum_{i=1}^{m} p_i^{(1)} \{\hat{\pi}^{[j]}(X_i, Z_i, W_i) - \bar{\pi}_n^{[j]}\} = 0\ (j = 1, \ldots, J),$$
$$\sum_{i=1}^{m} p_i^{(1)} \{W_i \hat{g}_\theta^{[k]}(X_i, Z_i, W_i) - \bar{g}_\theta^{w[k]}\} = 0\ (k = 1, \ldots, K),$$

where $\bar{\pi}_n^{[j]}$ and $\bar{g}_\theta^{w[k]}$ are defined as

$$\bar{\pi}_n^{[j]} = n^{-1} \sum_{i=1}^{n} \hat{\pi}^{[j]}(X_i, Z_i, W_i),\ \bar{g}_\theta^{w[k]} = n^{-1} \sum_{i=1}^{n} W_i \hat{g}_\theta^{[k]}(X_i, Z_i, W_i),$$

respectively. The first-step maximum empirical weights are denoted as

$$(\hat{p}_1^{(1)}, \ldots, \hat{p}_m^{(1)}) = (\hat{p}_1^{(1)}(\hat{\theta}_1), \ldots, \hat{p}_m^{(1)}(\hat{\theta}_1)).$$

Next, we define the maximum empirical likelihood weights in the second step as

$$\hat{\theta}_2 = \arg\max_{\theta} \arg\max_{p_1^{(2)},\ldots,p_N^{(2)}} \sum_{i=1}^{N} \log p_i^{(2)},$$

subject to

$$\sum_{i=1}^{N} p_i^{(2)} = 1, \ \sum_{i=1}^{N} p_i^{(2)} \left(1 - \delta_i W_i\right) \hat{C}_\theta^{[l]}(X_i) = 0 \quad (l = 1, \ldots, L).$$

The second-step maximum empirical weights are denoted as

$$(\hat{p}_1^{(2)}, \ldots, \hat{p}_m^{(2)}) = (\hat{p}_1^{(2)}(\hat{\theta}_2), \ldots, \hat{p}_m^{(2)}(\hat{\theta}_2)).$$

Then, we define the final maximum empirical likelihood estimator $\hat{\theta}_{\mathrm{EL1}}$ in Setting 1 as the unique solution to

$$\sum_{i=1}^{N} \hat{p}_i^{(2)} \hat{p}_i^{(1)} \delta_i R_i W_i U_\theta(X_i, Y_i) = 0. \tag{27}$$

Remark 3. The first step of the empirical weights described above is a generalization of the multiple-robust estimator in Han (2014). Indeed, $W_i = 1\,(i = 1, \ldots, n)$ yields the same empirical weights, implying that multiplication by W_i is required to adjust the sampling bias under informative sampling. The concept for the second step of the empirical weights comes from Qin, Zhang and Leung (2009)'s estimator $\hat{\theta}_{\mathrm{EL,Q}}$. Therefore, our empirical likelihood weights are obtained by combining the ideas of Han (2014) and Qin, Zhang and Leung (2009).

Remark 4. Our empirical likelihood estimator does not directly use $1/\hat{\pi}(x, z, w)$, unlike the method of moments estimator. Therefore, the finite-sample performance of the empirical likelihood estimator is better than that of the method of moments estimator when some $\hat{\pi}(x, z, w)$ can take on values near zero. See Han (2014) for additional details.

In Setting 2, we must modify the empirical weights in the second step and define the maximum empirical likelihood estimator as

$$\hat{V} = \arg\max_{V} \arg\max_{p_1^{(2)},\ldots,p_n^{(2)}} \sum_{i=1}^{n} \log p_i^{(2)} + (N - n)\log(1 - V),$$

subject to

$$\sum_{i=1}^{n} p_i^{(2)} = 1, \ \sum_{i=1}^{n} p_i^{(2)}(1/W_i - V) = 0.$$

Then, we define the empirical weights as $(\hat{p}_1^{(2)}, \ldots, \hat{p}_n^{(2)}) = (\hat{p}_1^{(2)}(\hat{V}), \ldots, \hat{p}_n^{(2)}(\hat{V}))$. Our maximum empirical likelihood estimator $\hat{\theta}_{\mathrm{EL2}}$ in Setting 2 is defined as the unique solution to

$$\sum_{i=1}^{N} \hat{p}_i^{(2)} \hat{p}_i^{(1)} \delta_i R_i U_\theta(X_i, Y_i) = 0. \tag{28}$$

In Setting 2, modeling C_θ and multiplying by the sampling weights in (28) are unnecessary because such information is already carried by the second term in the empirical likelihood. Then, our estimators $\hat{\theta}_{\mathrm{EL1}}$ and $\hat{\theta}_{\mathrm{EL2}}$ have desired multiple robustness.

Theorem 3.4. *In Settings 1 and 2, we assume Conditions 1–2 and Conditions 7–10 in Appendix. If each of the J models for the response mechanism, K models for the regression function, and L models for $C_\theta(x)$ include the correct model, then the two adaptive estimators $\hat{\theta}_{\mathrm{EL1}}$ and $\hat{\theta}_{\mathrm{EL2}}$ achieve the efficiency bound $\{E(S_{\mathrm{eff},j}^{\otimes 2})\}^{-1}$ for each Setting $j = 1, 2$. If at least one of the $J + K$ models for the response and outcome regression models is correctly specified, then $\hat{\theta}_{\mathrm{EL1}}$ and $\hat{\theta}_{\mathrm{EL2}}$ still have consistency.*

3.1.5 Optimal estimator using empirical likelihood in Setting 3

Finally, we propose the most efficient estimator in Setting 3 by extending the estimator in Setting 2. Recall that $\tau = \tau(P_1)$ is the target parameter in the external data, where P_1 is the distribution of external data. Suppose that an estimator $\tilde{\tau}$ for τ, its variance estimator $\tilde{\Sigma}_1$, and the sample size of the external data N_1 are available from the external source. Additionally, suppose that we can access summary statistics such as (i) Z-estimator (the solution to $E\{U_\tau(X, Z)\} = 0$), (ii) regression coefficients $E(X \mid Z; \tau)$, and (iii) conditional density $f(x \mid z; \tau)$. According to Condition 3, because there is no missingness for X and Z after sampling, the optimal estimator is obtained by solving $n^{-1} \sum_{i=1}^{N} \{\delta_i W_i D_\tau^*(X_i, Z_i) + (1 - \delta_i W_i) C_\tau^*(X_i)\} = 0$, where D_τ^* and C_τ^* are defined in Theorem 3.1 in Morikawa, Terada and Kim (2022) dependent on the target parameter τ. For example, if a Z-estimator is of interest, then $D_\theta^* = U_\tau(X, Z)$ and $C_\tau^* = E\{(W - 1)U_\tau(X, Z) \mid X\}/E(W - 1 \mid X)$.

Then, by using the optimal score function for τ, the maximum empirical likelihood estimator in Setting 3 in the second step is defined through the maximizer of (τ, V) in

$$\operatorname*{arg\,max}_{p_1^{(2)}, \ldots, p_n^{(2)}} \sum_{i=1}^{n} \log p_i^{(2)} + (N - n) \log(1 - V) \; - 2^{-1} N_1 (\tilde{\tau} - \tau)^\top \tilde{\Sigma}_1^{-1} (\tilde{\tau} - \tau)^\top,$$

subject to

$$\sum_{i=1}^{n} p_i^{(2)} = 1, \quad \sum_{i=1}^{n} p_i^{(2)} (1/W_i - V) = 0, \; \sum_{i=1}^{n} p_i^{(2)} D_\tau^*(X_i, Z_i) = 0.$$

We incorporate the information on $\tilde{\tau}$ into the likelihood to leverage our prior knowledge that $\tilde{\tau}$ is asymptotically normally distributed with mean τ and variance $\tilde{\Sigma}_1$, as discussed in Zhang et al. (2020). Furthermore, the third constraint above is essential for efficiently estimating τ using the internal data. Then, our final estimator $\hat{\theta}_{\mathrm{EL3}}$ is obtained by solving (28) with respect to θ.

Theorem 3.5. *In Setting 3, we assume Conditions 1–6, Conditions 7–9 and 11 in Appendix. If each of the J models for the response mechanism and K models for the regression function include the correct model, the adaptive estimator $\hat{\theta}_{\mathrm{EL3}}$ achieves the efficiency bound addressed in section 3.1.3. If at least one of the $J + K$ models for the response and outcome regression models is correctly specified, $\hat{\theta}_{\mathrm{EL3}}$ still has consistency.*

3.2 Future work: Nonignorable participation for non-probability survey samples

In section 2, we discussed the identifiability of nonignorable missing data. The proposed identification conditions are beneficial in settings where the data structure can be viewed within the framework of incomplete data analysis. In this section, as an illustrative example, we introduce the utilization of non-probability samples in survey sampling. First, we introduce the notation: the finite population $\mathcal{U} = \{1, 2, \ldots, N\}$ with population size N, auxiliary variables, $\boldsymbol{x}_i$, and the study variable, y_i. The finite population mean of the study variable is defined as $\mu_y = N^{-1}\sum_{i=1}^{N} y_i$. Let $\mathcal{S}_A$ be the set of n_A units for the non-probability survey sample and $\{(y_i, \boldsymbol{x}_i), i \in \mathcal{S}_A\}$ be the sample dataset; let $\mathcal{S}_B$ be the set of n_B units of an existing reference probability survey sample with the sample dataset represented by $\{(\boldsymbol{x}_i, d_i^B), i \in \mathcal{S}_B\}$, where the d_i^B's are the survey weights. Let $\delta_i = I(i \in \mathcal{S}_A)$ be the indicator variable for the participation of unit i in the non-probability sample $\mathcal{S}_A$, $i = 1, 2, \ldots, N$. In a probability sample, the inclusion probabilities d_i^B are known, but the study variable y is not observed. Conversely, in a non-probability sample, the inclusion probabilities are unknown, but the study variable y is observed.

In this setting, we want to estimate the parameter of the participation probabilities for the non-probability sample defined as

$$\pi_i = P(i \in \mathcal{S}_A \mid \boldsymbol{x}_i, y_i) = P(\delta_i = 1 \mid \boldsymbol{x}_i, y_i)\,,\ i = 1, 2, \ldots, N\,. \tag{29}$$

Note that it is consistent with the response mechanism in missing data analysis. Under the ignorable condition, Chen, Li and Wu (2020) propose the doubly robust estimator.

Hereafter, we consider the relaxation of ignorable assumption. Similarly to the technique used by Chen et al. (2020), we explain how to estimate it through nonignorable missing data analysis methods. Suppose that $\boldsymbol{x}$ is observed for all units in the finite population while y is only observed for the non-probability sample. In this case, the observations are $\{(R_i, R_i y_i, \boldsymbol{x}_i), i = 1, 2, \ldots, N\}$ that is same in missing data analysis described in section 2. In this case, we develop a parametric modeling method that is often used in practical applications. Of course, it is also possible to extend semiparametric or nonparametric models for nonignorable missingness, such as Kim and Yu (2011) and Morikawa, Kim and Kano (2017).

Recall that we solve the following mean score equation (2) to get the MLE estimator in the nonignorable missing setting,

$$\sum_{i=1}^{n}\left\{\delta_i \frac{\partial \log \pi(\boldsymbol{x}_i, y_i; \boldsymbol{\phi})}{\boldsymbol{\phi}} - (1-\delta_i)\frac{\int \partial\pi(\boldsymbol{x}_i, y; \boldsymbol{\phi})/\partial\boldsymbol{\phi}\cdot p(y \mid \boldsymbol{x})dy}{\int\{1-\pi(\boldsymbol{x}_i, y; \boldsymbol{\phi})\}p(y \mid \boldsymbol{x})dy}\right\} = 0.$$

Similarly to the technique used by Chen et al. (2020), the mean score equation (2) can be approximated as

$$\sum_{i\in\mathcal{S}_A}\left[\frac{\partial \log \pi(\boldsymbol{x}_i, y_i; \boldsymbol{\phi})}{\boldsymbol{\phi}} - \frac{\int \partial\pi(\boldsymbol{x}_i, y; \boldsymbol{\phi})/\partial\boldsymbol{\phi}\cdot p(y \mid \boldsymbol{x})dy}{\int\{1-\pi(\boldsymbol{x}_i, y; \boldsymbol{\phi})\}p(y \mid \boldsymbol{x})dy}\right] + \sum_{i\in\mathcal{S}_B} d_i^B \frac{\int \partial\pi(\boldsymbol{x}_i, y; \boldsymbol{\phi})/\partial\boldsymbol{\phi}\cdot p(y \mid \boldsymbol{x})dy}{\int\{1-\pi(\boldsymbol{x}_i, y; \boldsymbol{\phi})\}p(y \mid \boldsymbol{x})dy}. \tag{30}$$

Note that the identifiability with respect to the equation (2) is required because the equation (30) approximates the equation (2). To guarantee identifiability with respect to (2), one can directly use the conditions in section 2 which is the future work of this thesis.

4 Numerical experiment

In this section, we perform numerical experiments related to the identification conditions proposed in section 2.5 and 2.6, respectively.

4.1 Simulation of instrumental variables approach

We present the effects of identifiability in numerical experiments by comparing weak and strong identifiable models. We prepared four Scenarios S1–S4:

S1: (Outcome: Normal, Response: Logistic)
$[y \mid u, z, \delta = 1] \sim N(\kappa_0 + \kappa_1 u + \kappa_2 z, \sigma^2)$, $\text{logit}\{P(\delta = 1 \mid u, y; \boldsymbol{\alpha}, \beta)\} = \alpha_0 + \alpha_1 u + \beta y$, $u \sim N(0, 1^2)$, and $z \sim B(1, 0.5)$, where $(\kappa_0, \kappa_1, \sigma^2)^\top = (0.3, 0.4, 1/\sqrt{2}^2)^\top$ and $(\alpha_0, \alpha_1, \beta)^\top = (0.7, -0.2, 0.29)^\top$.

S2: (Outcome: Normal, Response: Cauchy)
$[y \mid u, z, \delta = 1] \sim N(\kappa_0 + \kappa_1 u + \kappa_2 z, \sigma^2)$, $P(\delta = 1 \mid u, y; \boldsymbol{\alpha}, \beta) = \Psi(\alpha_0 + \alpha_1 u + \beta y)$, $u \sim \text{Unif}(-1, 1)$, and $z \sim B(1, 0.7)$, where $(\kappa_0, \kappa_1, \sigma^2)^\top = (-0.36, 0.59, 1/\sqrt{2}^2)^\top$, $(\alpha_0, \alpha_1, \beta)^\top = (0.24, -0.1, 0.42)^\top$, and Ψ is the cumulative distribution function of the Cauchy distribution.

S3: (Outcome: Bernoulli, Response: Probit)
$[y \mid u, z, \delta = 1] \sim B(1, p(u, z; \boldsymbol{\kappa})\})$, $P(\delta = 1 \mid u, y; \boldsymbol{\alpha}, \beta) = \Psi(\alpha_0 + \alpha_1 u + \beta y)$, $u \sim N(0, 1^2)$, and $z \sim N(0, 1^2)$, where $p(u, z; \boldsymbol{\kappa}) = 1/\{1 + \exp(-\kappa_0 - \kappa_1 u - \kappa_2 z)$, $(\kappa_0, \kappa_1, \kappa_2)^\top = (-0.21, 3.8, 1.0)^\top$, $(\alpha_0, \alpha_1, \beta)^\top = (0.4, 0.39, 0.3)^\top$, and Ψ is the cumulative distribution function of the standard normal.

S4: (Outcome: Normal+nonlinear mean structure, Response: Cauchy or Logistic)
$[y \mid u, z, \delta = 1] \sim N(\mu(\boldsymbol{x}), 0.5^2)$, $P(\delta = 1 \mid u, y; \boldsymbol{\alpha}, \beta) = \Psi(\alpha_0 + \alpha_1 u + \beta y)$, $u \sim \text{Unif}(-1, 1)$, and $z \sim B(1, 0.5)$, where $\mu(\boldsymbol{x}) = z + \cos(2\pi u) + \exp(z + u)$, $(\alpha_0, \alpha_1, \beta)^\top = (0.1, -0.2, 0.3)^\top$, and Ψ is the cumulative distribution function of the Cauchy or Logistic distribution.

In S1 and S2, the strength of the identification can be adjusted by changing the parameter κ_2 because $\kappa_2 = 0$ indicates that the model it is unidentifiable by Example 2.1. On the other hand, we can verify that the models in S3 and S4 are identifiable by Theorem 2.8. For example, in S4, we can see that checking (C3) and (C4) is straightforward to the setting, while (C5) and (C6) hold from Example 2.9 and Proposition 2.11, respectively. From S3 and S4, we can confirm the successful inference even in the case of discrete outcome and complex mean structures, respectively.

We generated 1,000 independent Monte Carlo samples and computed two estimators for $E[y]$ and β with two methods: fractional imputation (FI) and complete case (CC) estimators, which use only completely observed data. The estimator for $E[y]$ is computed by the standard inverse probability weighting method with estimated response models (Riddles et al., 2016). We used correctly specified models for Scenarios S1–S3 but used nonparametric models for Scenario S4 because it is unrealistic to assume that the complicated mean structure is known. The R package ‘crs’ specialized in nonparametric spline regression on the mixture of categorical and continuous covariates (Nie and Racine, 2012) is used to estimate the respondents’ outcome model. Response model are estimated by using the method discussed in Section 2.2.

Bias, root mean squared error (RMSE), and coverage rate for 95% confidence intervals in S1–S4 are reported in Table 1. In all the Scenarios, CC estimators have a significant bias, and the coverage rates are far from 95%, while FI estimators work well when the model is surely identifiable. When κ_2 is small in S1 and S2, the performance of variance estimation with FI is poor, as expected, although that of point estimates is acceptable. The results in S4 indicate that the model is identifiable even if we use a nonparametric mean structure, and the estimates are almost the same between the two response models.

Table 1: Results of S1–S4: Bias, root mean square error (RMSE), and coverage rate (CR,%) with 95% confidence interval are reported. CC: complete case; FI: fractional imputation.

Scenario	Parameter	κ_2	Method	Bias	RMSE	CR
S1	$E[y]$	1.0	CC	0.053	0.066	73.5
			FI	0.000	0.043	95.4
		0.5	CC	0.039	0.053	80.9
			FI	-0.001	0.059	97.1
		0.1	CC	0.034	0.049	83.0
			FI	0.021	0.136	99.8
	β	1.0	FI	0.001	0.163	95.2
		0.5	FI	0.003	0.330	98.6
		0.1	FI	-0.146	0.865	100
S2	$E[y]$	1.0	CC	0.146	0.152	5.7
			FI	-0.004	0.051	94.8
		0.5	CC	0.130	0.136	7.7
			FI	-0.008	0.086	86.2
		0.1	CC	0.127	0.133	9.4
			FI	-0.007	0.105	92.4
	β	1.0	FI	0.008	0.148	95.4
		0.5	FI	0.044	0.365	100
		0.1	FI	0.033	0.448	100
S3	$E[y]$	–	CC	0.100	0.102	0.3
		–	FI	0.001	0.022	95.3
	β	–	FI	-0.023	0.279	95.0
S4	$E[y]$	–	CC(Logistic)	0.341	0.355	5.4
		–	FI(Logistic)	0.005	0.079	95.4
		–	CC(Cauchy)	0.296	0.312	10.7
		–	FI(Cauchy)	0.007	0.080	94.3
	β	–	FI(Logistic)	0.006	0.050	94.7
		–	FI(Cauchy)	0.011	0.063	93.8

4.2 Simulation of avoiding instrumental variables approach

In this section, we conduct numerical studies to evaluate the performance of the proposed FI method. We assume that the response mechanism follows a logistic distribution.

$$\text{logit}\,\{\text{pr}(\delta = 1 \mid x, y; \alpha, \beta)\} = \alpha_0 + \alpha_1 x + \beta y,$$

where x denotes a one-dimensional covariate. We conduct three scenarios, S1–S3 with varying outcome models, as follows:

- S1: The distribution of $[y \mid x, \delta = 1]$ is $N(\kappa_0 + \kappa_1 x + \kappa_2 x^2, \sigma^2)$, where $(\kappa_0, \kappa_1, \sigma^2)^\top = (0, 0.4, 1/\sqrt{2}^2)^\top$ and κ_2 (identifiability) are 0.1 (weak), 0.5 (moderate), and 1.0 (strong), the distribution of the covariate x is $N(0, 1^2)$, and the true parameter of the response mechanism is $(\alpha_0, \alpha_1, \beta)^\top = (0.68, 0.19, 0.24)^\top$.
- S2: The distribution of $[y \mid x, \delta = 1]$ follows a binomial distribution $B(1, 1/\{1 + \exp(-\kappa_0 - \kappa_1 x)\})$, where $(\kappa_0, \kappa_1)^\top = (-0.21, 5.9)^\top$, the distribution of the covariate $[x]$ is $N(0, 1^2)$, and the true parameter of the response mechanism is $(\alpha_0, \alpha_1, \beta)^\top = (0.7, 0.39, 0.39)^\top$.
- S3: The distribution of $[y \mid x, \delta = 1]$ follows a normal mixture distribution $0.35N(1 - 1.4x, 1/\sqrt{2}^2) + 0.65N(-1.5 - 0.5x + x^2, 1/\sqrt{2}^2)$, where the distribution of the covariate $[x]$ is $N(0, 1^2)$ and the true parameter of the response mechanism is $(\alpha_0, \alpha_1, \beta)^\top = (0.9, -0.26, 0.2)^\top$.

We generate $B = 1{,}000$ independent Monte Carlo samples with a sample size of $n = 500$. The average response rate is approximately 0.7 for all scenarios. We compare our proposed FI estimators with complete case (CC) estimators, which use only complete cases. For comparison, we consider two parameters: the expectation of the missing variable y, μ_y, and the response model coefficient β associated with the missing variable y.

Table 2 reports the bias, root-mean-square error (RMSE), and coverage rates with a 95% confidence interval. Although the naive CC estimators have large biases in all scenarios, the proposed FI estimators yield asymptotically unbiased estimates, except for non-identifiable situations.

Based on the discussion in Example 2.14, we demonstrate that the model in S1 is unidentifiable for $\kappa_2 = 0$. Moreover, as κ_2 tends to 0, the identifiability becomes weaker, hence yielding inaccurate estimates. For the estimations of μ_y, the FI estimators performed adequately in all scenarios, and the results for the coverage rate were acceptable. Regarding the estimations of β, the bias is still less in all scenarios. However, the RMSE in S2 takes large values owing to the lack of information on binary outcomes.

Table 2: Results of S1–S3: bias, root mean square error (RMSE), and coverage rate (CR,%) with 95% confidence interval are reported. CC, complete case; FI: fractional imputation.

Scenario	Parameter	κ_2	Method	Bias	RMSE	CR
S1	μ_y	1.0	CC	0.167	0.191	59.1
			FI	0.000	0.076	94.9
		0.5	CC	0.099	0.116	64.2
			FI	-0.002	0.057	95.5
		0.1	CC	0.077	0.089	60.8
			FI	-0.006	0.137	98.8
	β	1.0	FI	0.015	0.097	95.3
		0.5	FI	0.017	0.162	96.7
		0.1	FI	0.078	1.694	99.1
S2	μ_y	–	CC	0.073	0.078	21.5
		–	FI	0.002	0.026	95.1
	β	–	FI	0.018	0.508	95.6
S3	μ_y	–	CC	0.214	0.236	43.0
		–	FI	-0.001	0.102	94.9
	β	–	FI	0.013	0.165	95.3

5 Real data analysis

In this section, we perform real data analysis related to the identification conditions proposed in section 2.5 and 2.6, respectively.

5.1 AIDS Clinical Trials Group Study 175

We analyzed a dataset of 2139 HIV-positive patients enrolled in AIDS Clinical Trials Group Study 175 (ACTG175; Hammer, Katzenstein, Hughes, Gundacker, Schooley, Haubrich, Henry, Lederman, Phair, Niu et al. (1996)). In this analysis, we specify 532 patients for analysis who received zidovudine (ZDV) monotherapy. Let each y, x_1, and x_2 be the CD4 cell count at 96 ± 5 weeks, at the baseline, and at 20 ± 5 weeks, x_3 be the CD8 cell count at the baseline, and z be sex. The outcome was subject to missingness with a 60.34% observation rate, while all covariates were observed. To make estimation stable and easy, we standardized all the data. We expect that z (sex) is a reasonable choice for an instrument variable because the information is a biological value, which affects the value of CD4, but has little effect on the response probability.

Patients who are suffering from a mild illness of HIV tend to have higher CD4 cell count; thus, one may consider that missingness of the outcome relates to serious conditions and may expect that the missing value of the outcome would be a lower CD4 cell count than the respondent. We therefore considered five different MNAR response models:

$$P(\delta = 1 \mid x_1, x_2, x_3, y) = \Psi(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \beta y),$$

where Ψ represents either the Logistic function or the distribution functions of the Cauchy or t distribution with degrees of freedom $v\,(= 2, 5, 10)$. Theorem 2.8 and Proposition 2.11

ensure that all the models with these five response models are identifiable, even when the instrumental variable z is discrete. From the above conjecture on missing values, the sign of β is expected to be negative. We assumed that the respondent's outcome is a normal distribution with a nonparametric mean structure and estimated by the 'crs' R package as considered in Scenario S4 in Section 4. The residual plots shown in Figure 3 and the computed R^2-value (= 0.453) signify the assumed distribution on the respondents' outcome fit well. Table 3 reports the estimated parameters and their estimated standard errors calculated by 1,000 bootstrap samples. The results of the five response models were almost similar. This suggests that the response mechanism is robust to the choice of response models. Although we cannot determine whether it is MNAR or MAR because the estimated standard error for β is large, the point estimate is negative, as we expected. This result is consistent with the result in Zhao, Wang and Shao (2021).

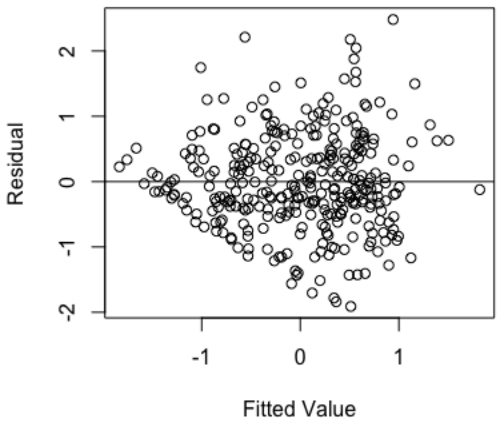

Figure 2: Residual plots of respondents' outcome in ACTG175 data.

Table 3: Estimated parameters: Estimates and standard errors for the target parameters are reported. Logistic and Cauchy are Fractional Imputation using Logistic and Cauchy distributions for the response mechanism. T_v: t distribution function with degrees of freedom $v\,(= 2, 5, 10)$.

Parameter	Model	Estimate	SE	Parameter	Model	Estimate	SE
	Logistic	0.464	0.104		Logistic	0.125	0.156
	Cauchy	0.417	0.260		Cauchy	0.108	0.139
α_0	T_2	0.341	0.081	α_1	T_2	0.091	0.113
	T_5	0.306	0.069		T_5	0.082	0.102
	T_{10}	0.295	0.066		T_{10}	0.080	0.099
	Logistic	0.255	0.192		Logistic	0.093	0.107
	Cauchy	0.244	0.207		Cauchy	0.083	0.097
α_2	T_2	0.196	0.148	α_3	T_2	0.069	0.079
	T_5	0.169	0.126		T_5	0.062	0.070
	T_{10}	0.160	0.120		T_{10}	0.060	0.068
	Logistic	-0.032	0.314		Logistic	276.70	13.476
	Cauchy	-0.030	0.387		Cauchy	276.51	14.107
β	T_2	-0.027	0.235	$E[y]$	T_2	276.57	13.437
	T_5	-0.021	0.203		T_5	276.61	13.271
	T_{10}	-0.019	0.194		T_{10}	276.63	13.217

5.2 Election Poll Data

We apply the proposed method to opinion poll data collected to predict the 2022 South Korean presidential election. Specifically, data were obtained from a telephone survey with a response rate of 8.96% from 896 individuals among 10,000 potential voters. The dataset includes respondents' voting intentions regarding electoral participation, voting preferences, and demographic information, such as sex, area, and age.

The respondents' voting preferences, denoted by y, were categorised into binary responses: 1 for the candidate of the ruling party and 0 for other candidates. Additionally we redefine the intention of electoral participation, denoted by δ, such that it equals 1 if a respondent is likely to participate in the election, and 0 otherwise. We assume that we do not observe y, whose voting intention, δ, is 0. Among the 896 respondents, we observe y for 841(93.9%) respondents and do not observe y for the remaining 55(6.1%) respondents. Moreover, we implicitly assume a nonignorable missing mechanism, such that voting intention is closely related to the respondents' voting preferences.

The covariate *age*(≥ 19) is continuous, *sex* is a binary non-ordered categorical variable which represents male status or not, and *area* is another non-ordered categorical variable with three merged administrative districts in South Korea. The "true" averaged voting preference rate of $E(y)$ is $\hat{\theta}_n = 0.300(= 269/896)$ calculated using the complete data, and the naive average voting rate is $\hat{\theta}_{naive} = 0.317(= 267/841)$ calculated using the observed data with $\delta = 1$.

To apply our proposed method, we define a non-ordered categorical variable z with six values that integrate the *sex* and *area* variables. Subsequently, we assume a binomial

Table 4: Results of the analysis of the 2022 South Korean Presidential Election data: estimates and 95% confidence intervals (CI) for the target parameters reported.

Parameter	Estimate	95% CI	Parameter	Estimate	95% CI
α_{01}	2.865	(1.659, 4.072)	α_{11}	0.489	(-0.021, 1.001)
α_{02}	3.304	(1.673, 4.934)	α_{12}	0.308	(-0.707, 1.324)
α_{03}	2.616	(1.651, 3.580)	α_{13}	0.535	(0.018, 1.051)
α_{04}	2.860	(1.750, 3.971)	α_{14}	-0.266	(-0.799, 0.267)
α_{05}	3.764	(1.323, 6.205)	α_{15}	0.310	(-2.078, 2.698)
α_{06}	4.848	(-0.467, 10.165)	α_{16}	-1.817	(-6.722, 3.086)
β	-0.461	(-2.921, 1.999)	$E[Y]$	0.321	(0.292, 0.350)

distribution $B(1, p_1(x, z))$ as the outcome model, where

$$\text{logit}\{p_1(x, z)\} = \sum_{d=1}^{6} I(z = d) \sum_{l=0}^{4} \kappa_{ld} x^l,$$

and κ_{ld} is the coefficient of the logistic regression for the l-th power of the covariate for $z = d$, and x represents the standardisation of *age*. For each value of z, we select the most suitable logistic regression model by AIC in a stepwise algorithm among $2^5 - 1$ models. The resulting logistic regression functions vary for each categorical variable. For example, the most suitable model is $-0.12 - 1.09x - 0.69x^2 + 0.55x^3$ for $z = 1$, and $-0.68 - 0.85x - 0.66x^2$ for $z = 6$.

We assumed that the response model

$$\text{logit}\,\{\text{pr}(\delta = 1 \mid x, z, y; \alpha, \beta)\} = \sum_{d=1}^{6} I(z = d)(\alpha_{0d} + \alpha_{1d} x) + \beta y,$$

and further estimate $(\alpha^\top, \beta)^\top$ and the mean voting preference using the proposed FI method. As explained in Example 2.17, this model can be identified without using instrumental variables.

Table 4 presents the point estimates and 95% confidence intervals of the model parameters and $E[y]$. The 95% confidence intervals for $z = 5$ and $z = 6$ are wider than the others, owing to the smaller sample size. We can assert that the proposed FI method performs well because the confidence interval of $E[y]$ contains the "true" average voting rate $\hat{\theta}_n = 0.300$. Although the point estimator of β significantly deviates from 0, we cannot determine whether it is MAR or MNAR because the 95% confidence interval contains 0 in this real data application.

5.3 National Supported Work Data

In the second application, we use publicly available data collected to evaluate the National supported work (NSW) demonstration project Lalonde (1986). The response indicator δ denotes a treatment indicator, and the covariates of the dataset are *age, education, black, nodegree*, where *age* and *education* are continuous, and *black* and *nodegree* are non-ordered categorical binary response variables that represent whether the race of a participant is black and they do not take degree or not, respectively. The integration of *black* and *nodegree* defines a new non-ordered categorical vairble z with four values. The covariates x_1 and x_2 represent the standardisation of *age* and *education*, respectively. We

Table 5: Results of the analysis of the National Supported Work data: estimates and 95% confidence intervals (CI) for the target parameters reported.

Parameter	Estimate	95% CI	Parameter	Estimate	95% CI
α_{01}	12.47	(-7.72, 32.67)	α_{11}	0.01	(-0.19, 0.22)
α_{02}	11.94	(-6.16, 30.05)	α_{12}	0.14	(-0.80, 1.09)
α_{03}	13.01	(-6.81, 32.83)	α_{13}	0.85	(0.23, 1.47)
α_{04}	14.48	(-16.11, 45.08)	α_{14}	0.24	(-1.35, 1.84)
α_{21}	-0.07	(-0.50, 0.36)	β	-1.42	(-3.59, 0.73)
α_{22}	1.15	(-1.19, 3.50)	$E[Y]$	9.03	(8.50, 9.56)
α_{23}	0.57	(-0.11, 1.27)			
α_{24}	-1.50	(-11.48, 8.47)			

also define the outcome value y by the logarithmic transformation of earnings in 1978 using any non-zero. The response rate was approximately 43.7% for the 526 experimental participants.

We selected the most suitable model using the AIC and confirmed the identifiability of the model. As explained in Theorem 2.20, the most suitable model must satisfy (C2) and (C4): Hence, we verify how far the most suitable model and the model that does not satisfy (C2) or (C4) are separated by the AIC. For each value of z, we selected the most suitable normal mixture model using the adaptive lasso. We further determined the number of components by comparing the AIC of each component (Städler, Bühlmann and Van De Geer, 2010). We set the maximum number of the mixture components among the candidate models to three with basis functions $(x_1, x_2, x_1^2, x_2^2, x_1^3, x_2^3, x_1x_2)$ as covariates, whereas we set the number of mixture components of $z = 4$ to 1 because the sample size was small. The resulting regression functions vary for each categorical variable. For example, the most suitable model for $z = 1$ and $z = 3$ is

$$
\begin{aligned}
z = 1 :\ & 0.25N(7.13 - 0.68x_1 - 0.55x_2 + 0.06x_1^3, 0.90^2) \\
& + 0.33N(9.21 - 0.02x_2^2, 0.20^2) + 0.42N(8.86 + 0.44x_2 + 0.16x_1x_2, 0.68^2); \\
z = 3 :\ & 0.54N(9.62 + 0.07x_1^2 - 0.27x_2^2 + 0.88x_1^3 + 0.44x_1x_2, 0.21^2) \\
& + 0.45N(8.03 - 0.54x_2 + 0.08x_1^2 - 0.20x_2^2, 0.48^2).
\end{aligned}
$$

As explained in Theorem 2.20, this model is identifiable. For $z = 2$ and $z = 3$, the largest AIC models in unidentifiable models that do not satisfy (C2) or (C4) of Theoerem2.20 are 131.58 and 90.23. However, the most suitable models are 119.76 and 79.48, respectively. The naive average of the outcome is $\hat{\mu}_{naive} = 8.52$, calculated using only the respondents' data. Figure 3 reports the residual plots for each categorical variable, which show the goodness of fit of the normal mixture. We assume that the response model

$$
\operatorname{logit}\{\operatorname{pr}(\delta = 1 \mid x_1, x_2, z, y; \alpha, \beta)\} = \sum_{d=1}^{4} I(z = d)(\alpha_{0d} + \alpha_{1d}x_1 + \alpha_{2d}x_2) + \beta y,
$$

and further estimate $(\alpha^\top, \beta)^\top$ and the mean of y using the proposed FI method.

Table 5 reports the point estimate and 95% confidence intervals for each target parameter. A smaller sample size of $z = 4$ widens the 95% confidence intervals of α_{24}. Although the point estimate of β significantly deviates from 0, we cannot determine whether it is MAR or MNAR because the 95% confidence interval contains 0.

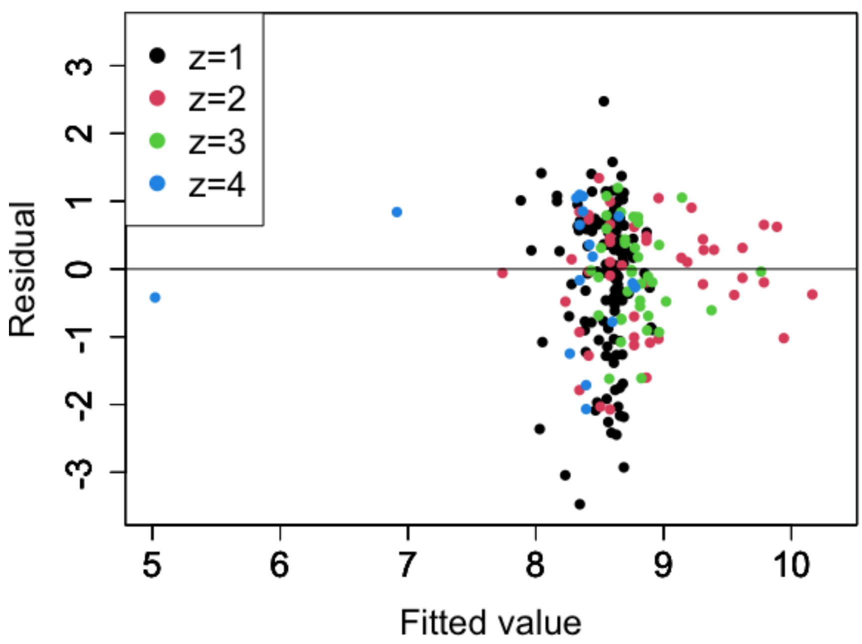

Figure 3: Residual plots for each z.

6 Conclusion

In this paper, we first proposed two new types of identification conditions for models using respondents' outcome and response models. Next, we developed efficient multiple robust estimation methods under informative sampling. Additionally, the integration of external summary statistics can also be easily handled using the proposed empirical likelihood approach.

Regarding the identification condition proposed in section 2.5, while our method requires the specification of the two models, these models can be highly flexible with the help of an instrument. As considered in Scenario S4 in section 4.1, the mean function in the respondents' outcome model can be nonparametric, and the response model can be any strictly monotone function, other than Logistic models. Our condition guarantees model identifiability even when instruments are categorical, which is not addressed by previous conditions. However, our method has certain limitations. First, respondents' outcome models need to satisfy the monotone-likelihood property in Condition (C5). For example, mixture models fall outside the scope of our framework. Second, the specification of instruments is necessary in advance. While recent studies, such as Zhao et al. (2021), have explored methods for identifying instruments, a universally accepted gold standard has yet to be established. In section 2.6, we propose sufficient conditions for model identifiability using a generalized linear model and a logistic response mechanism. We futher extend the outcome model to accommodate a mixture of distributions belonging to the exponential family and discuss the model identifiability of normal mixture models. However, the proposed FI method has notable limitations: it relies on the unverifiable assumption that if the response mechanism is nonignorable, it must be additive, meaning that the missing data y is related to the response only through the additive term βy. If these assumptions deviate significantly from reality, the proposed method may produce highly biased estimates. Although we restrict the response mechanism to a logistic distribution, it may be possible to utilize other distributions such as the Tobit and Robit models (Liu, 2004). However, these alternatives face additional challenges: the integral involved in the observed likelihood cannot typically be represented explicitly, and in some cases, this integral may diverge to infinity. Therefore, careful investigation is necessary when employing alternative response mechanisms. The proposed FI method can be replaced by multiple imputation, which is a popular method of missing data analysis (Rubin, 1978). Rubin's variance formula simplifies the calculation of the asymptotic variance of estimators. However, the congeniality condition requires further discussion to guarantee the applicability of Rubin's variance formula, which represents the scope of future work.

In the discussion of the efficient multiple robust estimations presented in section 3, the estimator achieves the semiparametric efficient bound as long as the model candidates include at least one correct model. However, there are also inherent limitations to this approach. Firstly, Condition 2, which assumes the ignorability of the response mechanism, can be restrictive in practical applications. Therefore, enhancing the methodology to accommodate nonignorable response mechanisms is critical for further development. Second, our approach assumes that external information is fully observed and excludes any parameters related to the outcome variable (y) from the current framework. Extending the method to incorporate such parameters represents an important direction for future research. Finally, the methods introduced in section 3.2 are still at a conceptual stage and remain a challenge for future investigation.

References

Baraldi, A. N., and Enders, C. K. (2010), "An introduction to modern missing data analyses," *Journal of school psychology*, **48**(1), 5–37.

Beppu, K., Choi, J., Morikawa, K., and Im, J. (2024), "Identification enhanced generalised linear model estimation with nonignorable missing outcomes," *arXiv preprint arXiv:2204.10508*, .

Berg, E., Kim, J., and Skinner, C. (2016), "Imputation under informative sampling," *J. Surv. Stat. Methodol.*, **4**, 436–462.

Binder, D. A. (1983), "On the variances of asymptotically normal estimators from complex surveys," *Int. Statist. Rev.*, **51**, 279–292.

Chang, T., and Kott, P. S. (2008), "Using calibration weighting to adjust for nonresponse under a plausible model," *Biometrika*, **95**(3), 555–571.

Chatterjee, N., Chen, Y.-H., Maas, P., and Carroll, R. (2016), "Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources," *J. Am. Statist. Assoc.*, **111**, 107–117.

Chen, S., and Haziza, D. (2017), "Multiply robust imputation procedures for the treatment of item nonresponse in surveys," *Biometrika*, **104**, 439–453.

Chen, Y., Li, P., and Wu, C. (2020), "Doubly robust inference with nonprobability survey samples," *Journal of the American Statistical Association*, **115**(532), 2011–2021.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018), "Double/debiased machine learning for treatment and structural parameters," *Econom. J.*, **21**(1), C1–C68.

Cui, X., Guo, J., and Yang, G. (2017), "On the identifiability and estimation of generalized linear models with parametric nonignorable missing data mechanism," *Computational Statistics & Data Analysis*, **107**, 64–80.

D'Haultfœuille, X. (2010), "A new instrumental method for dealing with endogenous selection," *Journal of Econometrics*, **154**(1), 1–15.

D'Haultfœuille, X. (2011), "On the completeness condition in nonparametric instrumental problems," *Econometric Theory*, **27**(3), 460–471.

Franks, A. M., Airoldib, E. M., and Rubin, D. B. (2020), "Nonstandard conditionally specified models fornonignorable missing data," *Proceedings of the National Academy of Sciences*, **117**(32), 19045–19053.

Hájek, J. (1970), "A characterization of limiting distributions of regular estimates," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **14**(4), 323–330.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M. et al. (1996), "A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter," *New England Journal of Medicine*, **335**(15), 1081–1090.

Han, P. (2014), "Multiply Robust Estimation in Regression Analysis With Missing Data," *J. Am. Statist. Assoc.*, **109**, 1159–1173.

Hazewinkel, A.-D., Bowden, J., Wade, K. H., Palmer, T., Wiles, N. J., and Tilling, K. (2022), "Sensitivity to missing not at random dropout in clinical trials: Use and interpretation of the trimmed means estimator," *Statistics in Medicine*, **41**(8), 1462–1481.

Hicks, S. C., Townes, F. W., Teng, M., and Irizarry, R. A. (2018), "Missing data and technical variability in single-cell RNA-sequencing experiments," *Biostatistics*, **19**(4), 562–578.

Hu, W., Wang, R., Li, W., and Miao, W. (2022), "Paradoxes and resolutions for semiparametric fusion of individual and summary data," *arXiv preprint arXiv:2210.00200*, .

Ibrahim, J., Chen, M.-H., Lipsitz, S., and Herring, A. (2005), "Missing-Data Methods for Generalized Linear Models," *Journal of the American Statistical Association*, **100**, 332–347.

Ibrahim, J. G., Lipsitz, S. R., and Horton, N. (2001), "Using auxiliary data for parameter estimation with non-ignorably missing outcomes," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **50**(3), 361–373.

Im, J., Cho, I.-H., and Kim, J.-K. (2018), "FHDI: An R Package for Fractional Hot Deck Imputation," *The R Journal*, pp. 140–154.

Imbens, G. W., and Rubin, D. B. (2015), *Causal inference in statistics, social, and biomedical sciences* Cambridge University Press.

Inagaki, N. (1970), "On the limiting distribution of a sequence of estimators with uniformity property," *Annals of the Institute of Statistical Mathematics*, **22**(1), 1–13.

Kim, J. K. (2011), "Parametric fractional imputation for missing data analysis," *Biometrika*, **98**(1), 119–132.

Kim, J. K., and Haziza, D. (2014), "Doubly Robust Inference With Missing Data in Survey Sampling," *Statist. Sin.*, **24**, 375–394.

Kim, J. K., and Shao, J. (2013), *Statistical Methods for Handling Incomplete Data* Chapman & Hall/CRC.

Kim, J. K., and Shao, J. (2021), *Statistical methods for handling incomplete data* CRC press.

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**(493), 157–165.

Kundu, P., Tang, R., and Chatterjee, N. (2019), "Generalized meta-analysis for multiple regression models across studies with disparate covariate information," *Biometrika*, **106**, 567–585.

Lalonde, R. (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *The American Economic Review*, **76**(4), 604–620.

Leurent, B., Gomes, M., Faria, R., Morris, S., Grieve, R., and Carpenter, J. R. (2018), "Sensitivity analysis for not-at-random missing data in trial-based cost-effectiveness analysis: a tutorial," *Pharmacoeconomics*, **36**(8), 889–901.

Li, M., Ma, Y., and Zhao, J. (2021), "Efficient estimation in a partially specified nonignorable propensity score model," *Computational Statistics & Data Analysis*, p. 107322.

Li, P., Qin, J., and Liu, Y. (2023), "Instability of inverse probability weighting methods and a remedy for nonignorable missing data," *Biometrics*, **79**(4), 3215–3226.

Little, R. (2003), "Bayesian methods for unit and item nonresponse," In Chambers, R.L. and Skinner, C. J. Analysis of Survey Data. Wiley*, Chichester*, pp. 289–306.

Little, R. J., and Rubin, D. B. (2019), *Statistical analysis with missing data*, Vol. 793 John Wiley & Sons.

Liu, C. (2004), "Robit regression: a simple robust alternative to logistic and probit regression," *Applied Bayesian Modeling and Casual Inference from Incomplete-Data Perspectives*, pp. 227–238.

Liu, Y., Li, P., and Qin, J. (2022), "Full-semiparametric-likelihood-based inference for non-ignorable missing data," *Statistica Sinica*, **32**(1), 271–292.

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**(516), 1673–1683.

Miao, W., Liu, L., Tchetgen, E. T., and Geng, Z. (2019), "Identification, doubly robust estimation, and semiparametric efficiency theory of nonignorable missing data with a shadow variable," *arXiv preprint arXiv:1509.02556*, .

Miao, W., and Tchetgen, E. T. (2018), "Identification and inference with nonignorable missing covariate data," *Statistica Sinica*, **28**(4), 2049.

Miao, W., and Tchetgen Tchetgen, E. J. (2016), "On varieties of doubly robust estimators under missingness not at random with a shadow variable," *Biometrika*, **103**(2), 475–482.

Morikawa, K., Beppu, K., and Aida, W. (2023), "Efficient Multiple-Robust Estimation for Nonresponse Data Under Informative Sampling," *arXiv preprint arXiv:2311.06719*, .

Morikawa, K., and Kim, J. K. (2021), "Semiparametric optimal estimation with nonignorable nonresponse data," *The Annals of Statistics*, **49**(5), 2991–3014.

Morikawa, K., Kim, J. K., and Kano, Y. (2017), "Semiparametric maximum likelihood estimation with data missing not at random," *Canadian Journal of Statistics*, **45**(4), 393–409.

Morikawa, K., Terada, Y., and Kim, J. K. (2022), "Semiparametric adaptive estimation under informative sampling," , . arXiv:2208.06039.

Nie, Z., and Racine, J. S. (2012), "The crs Package: Nonparametric Regression Splines for Continuous and Categorical Predictors.," *R Journal*, **4**(2).

Parent, M. C. (2013), "Handling item-level missing data: Simpler is just as good," *The Counseling Psychologist*, **41**(4), 568–600.

Peng, J., Hahn, J., and Huang, K.-W. (2023), "Handling missing values in information systems research: A review of methods and assumptions," *Information Systems Research*, **34**(1), 5–26.

Pfeffermann, D. (1993), "The role of sampling weights when modeling survey data," *Int. Statist. Rev.*, **61**, 317–337.

Pfeffermann, D., and Sverchkov, M. (2009), "Inference under informative sampling," *Handbook of Statistics, vol. 29B. Amsterdam: Elsevier*, pp. 455–487.

Qin, J., and Lawless, J. (1994), "Empirical likelihood and general estimating equations," *Ann. Statist.*, **22**(1), 300–325.

Qin, J., Leung, D., and Shao, J. (2002), "Estimation with survey data under nonignorable nonresponse or informative sampling," *J. Am. Statist. Assoc.*, **97**(457), 193–200.

Qin, J., Zhang, B., and Leung, H. Y. (2009), "Empirical Likelihood in Missing Data Problems," *J. Am. Statist. Assoc.*, pp. 1492–1503.

Riddles, M. K., Kim, J. K., and Im, J. (2016), "A propensity-score-adjustment method for nonignorable nonresponse," *Journal of Survey Statistics and Methodology*, **4**(2), 215–245.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994*a*), "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American statistical Association*, **89**(427), 846–866.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994*b*), "Estimation of regression coefficients when some regressors are not always observed," *J. Am. Statist. Assoc.*, **89**, 846–866.

Rubin, D. B. (1976), "Inference and missing data," *Biometrika*, **63**(3), 581–592.

Rubin, D. B. (1978), Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse,, in *Proceedings of the survey research methods section of the American Statistical Association*, Vol. 1, American Statistical Association, pp. 20–34.

Särndal, C. E. (1992), "Method for estimating the precision of survey estimates when imputation has been used," *Survey Methodol.*, **18**, 241–252.

Sauerbrei, W., Abrahamowicz, M., Altman, D. G., le Cessie, S., Carpenter, J., and initiative, S. (2014), "STRengthening analytical thinking for observational studies: the STRATOS initiative," *Statistics in medicine*, **33**(30), 5413–5432.

Shao, J., and Wang, L. (2016), "Semiparametric inverse propensity weighting for nonignorable missing data," *Biometrika*, **103**(1), 175–187.

Shetty, S., Ma, Y., and Zhao, J. (2021), "Avoid Estimating the Unknown Function in a Semiparametric Nonignorable Propensity Model," *arXiv preprint arXiv:2108.04966*, .

Städler, N., Bühlmann, P., and Van De Geer, S. (2010), "l_1-penalization for mixture regression models," *Test*, **19**(2), 209–256.

Tang, G., Little, R. J., and Raghunathan, T. E. (2003), "Analysis of multivariate missing data with nonignorable nonresponse," *Biometrika*, **90**(4), 747–764.

Wang, H., Lu, Z., and Liu, Y. (2021), "Score test for missing at random or not," *arXiv preprint arXiv:2105.12921*, .

Wang, S., Shao, J., and Kim, J. K. (2014), "An instrumental variable approach for identification and estimation with nonignorable nonresponse," *Statistica Sinica*, **24**, 1097–1116.

Yang, S., and Kim, J. K. (2020), "Statistical data integration in survey sampling: A review," *Japanese Journal of Statistics and Data Science*, **3**, 625–650.

Yang, S., Wang, L., and Ding, P. (2019), "Causal inference with confounders missing not at random," *Biometrika*, **106**(4), 875–888.

Zhang, H., Deng, L., Schiffman, M., Qin, J., and Yu, K. (2020), "Generalized integration model for improved statistical inference by leveraging external summary data," *Biometrika*, **107**(3), 689–703.

Zhao, J., and Ma, Y. (2022), "A versatile estimation procedure without estimating the nonignorable missingness mechanism," *Journal of the American Statistical Association*, **117**(540), 1916–1930.

Zhao, P., Wang, L., and Shao, J. (2021), "Sufficient dimension reduction and instrument search for data with nonignorable nonresponse," *Bernoulli*, **27**, 930–945.

Zou, H. (2006), "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, **101**(476), 1418–1429.

Appendix

A Technical Proofs of Section 2

We first provide a technical result to prove Theorem 2.7.

Lemma A.1. *Let a, b, and c be any positive real numbers. Assume that r_1 and r_2 are positive real numbers satisfying*

$$-\frac{ab}{a+b} < \frac{r_1^2 - r_2^2}{r_1^2 r_2^2} < c. \tag{31}$$

Then, there exist $0 < \pi_j^{(k)} < 1$ $(j = 1, 2, 3; k = 1, 2)$ such that

$$\sum_{j=1}^{3} \pi_j^{(1)} = r_1^2, \quad \sum_{j=1}^{3} \pi_j^{(2)} = r_2^2, \tag{32}$$

and

$$\frac{1}{\pi_1^{(1)}} - \frac{1}{\pi_1^{(2)}} = a, \quad \frac{1}{\pi_2^{(1)}} - \frac{1}{\pi_2^{(2)}} = b, \quad \frac{1}{\pi_3^{(1)}} - \frac{1}{\pi_3^{(2)}} = -c. \tag{33}$$

Proof of Lemma A.1. By using a polar coordinate system, we transform $\pi_j^{(k)}$ $(j = 1, 2, 3; k = 1, 2)$ into

$$\begin{aligned}
\left(\sqrt{\pi_1^{(1)}}, \sqrt{\pi_2^{(1)}}, \sqrt{\pi_3^{(1)}}\right) &= r_1(\sin\phi_1 \cos\phi_2, \sin\phi_1 \sin\phi_2, \cos\phi_1),\\
\left(\sqrt{\pi_1^{(2)}}, \sqrt{\pi_2^{(2)}}, \sqrt{\pi_3^{(2)}}\right) &= r_2(\sin\psi_1 \cos\psi_2, \sin\psi_1 \sin\psi_2, \cos\psi_1),
\end{aligned}$$

where $0 < \phi_1, \phi_2, \psi_1, \psi_2 < \pi/2$ to ensure $\pi_j^{(k)}$ $(j = 1, 2, 3; k = 1, 2)$ satisfy (32). It follows from (33) and double-angular formulas that we have

$$\begin{aligned}
&r_1^2(1-\omega_1)(1+\omega_2) - r_2^2(1-\omega_3)(1+\omega_4)\\
&= -\frac{ar_1^2 r_2^2}{4}(1-\omega_1)(1+\omega_2)(1-\omega_3)(1+\omega_4),
\end{aligned} \tag{34}$$

$$\begin{aligned}
&r_1^2(1-\omega_1)(1-\omega_2) - r_2^2(1-\omega_3)(1-\omega_4)\\
&= -\frac{br_1^2 r_2^2}{4}(1-\omega_1)(1-\omega_2)(1-\omega_3)(1-\omega_4),
\end{aligned} \tag{35}$$

$$r_1^2(1+\omega_1) - r_2^2(1+\omega_3) = \frac{cr_1^2 r_2^2}{2}(1+\omega_1)(1+\omega_3), \tag{36}$$

where $\omega_1 = \cos 2\phi_1, \omega_2 = \cos 2\phi_2, \omega_3 = \cos 2\psi_1$, and $\omega_4 = \cos 2\psi_2$. Setting $\omega_2 = \omega_4$ and equations (34) and (35) yield

$$\begin{aligned}
r_1^2(1-\omega_1) - r_2^2(1-\omega_3) &= -\frac{ar_1^2 r_2^2}{4}(1-\omega_1)(1+\omega_2)(1-\omega_3),\\
r_1^2(1-\omega_1) - r_2^2(1-\omega_3) &= -\frac{br_1^2 r_2^2}{4}(1-\omega_1)(1-\omega_2)(1-\omega_3).
\end{aligned}$$

Fixing $\omega_2 = 1 - 2a/(a+b)$ reduces the above equations to the one common equation

$$r_1^2(1-\omega_1) - r_2^2(1-\omega_3) = -\frac{r_1^2 r_2^2 ab}{2(a+b)}(1-\omega_1)(1-\omega_3), \tag{37}$$

maintaing the condition $-1 < \omega_2 < 1$. It remains to show that there exists $-1 < \omega_3 < 1$ satisfying (36) and (37). Solving the equation (37) with respect to ω_1, we have

$$\omega_1 = 1 - \frac{r_2^2(1-\omega_3)}{r_1^2 + r_1^2 r_2^2 ab(1-\omega_3)/\{2(a+b)\}}. \tag{38}$$

Substituting (38) into (36) leads to the following quadratic equation with respect to ω_3:

$$f(\omega_3) = \left(\frac{r_1^2 r_2^4 ab + c r_1^4 r_2^4 ab}{2(a+b)} - \frac{c r_1^2 r_2^4}{2}\right)\omega_3^2 - \left(\frac{r_1^4 r_2^2 ab}{a+b} + c r_1^4 r_2^2\right)\omega_3$$
$$+ \left(\frac{r_1^2 r_2^2 ab\,(2r_1^2 - r_2^2 - c r_1^2 r_2^2)}{2(a+b)} + \frac{c r_1^2 r_2^4}{2} + 2r_1^4 - 2r_1^2 r_2^2 - c r_1 - 4r_2^2\right) = 0.$$

It follows from (31) that

$$f(1) = r_1^2\left(2r_1^2 - 2r_2^2 - 2c r_1^2 r_2^2\right) < 0, \quad f(-1) = 2r_1^2\left(r_1^2 - r_2^2 + \frac{r_1^2 r_2^2 ab}{a+b}\right) > 0,$$

which implies that there is at least one solution of ω_3 to the equation $f(\omega_3) = 0$ in the open interval $(-1, 1)$.

□

Finally, we prove Theorem 2.7 with the help of Lemma A.1.

Proof of Theorem 2.7. Without loss of generality, we set the value of $\boldsymbol{u}$ to a fixed vector because the following proof holds for each $\boldsymbol{u}$. Let the categorical variables y and z take values in $\{1, 2, \dots, p\}$ and $\{1, 2, \dots, q\}$, respectively. We show that model identifiability implies the completeness condition (C2) by individually addressing three cases: (i) $p = 2$, (ii) $p = 3$, and (iii) $p \geq 4$ because "if" part has been already established by Lemma 2.5.

When $p = 2$, condition (C1) results in the rank of a $q \times 2$ matrix, composed of $p(y = j \mid \delta = 1, z = i)$ in its (i, j)-th element ($i = 1, 2; j = 1 \dots, q$), being 2. Hence, identifiable models always satisfy the completeness condition (C2).

For cases where $p \geq 3$, we must show that the model becomes unidentifiable when the completeness condition is violated. The breach of the completeness condition indicates the existence of a non-zero vector $(h_1, \dots, h_p)$ such that for $z = 1, \dots, q$, we have

$$E[h_y \mid \delta = 1, z] = \sum_{y=1}^{p} h_y p(y \mid \delta = 1, z) = 0. \tag{39}$$

The elements in $(h_1, \cdots, h_p)$ do not all share the same sign, and multiplying this vector by any constant does not affect the above equation. Recall that the model's unidentifiability implies that $\pi_y^{(1)} \neq \pi_y^{(2)}$ exists for some $y \in \{1, \dots, p\}$, satisfying $\sum_{y=1}^{p} p(y \mid \delta = 1, z)/\pi_y^{(1)} = \sum_{y=1}^{p} p(y \mid \delta = 1, z)/\pi_y^{(2)}$. We now construct an unidentifiable model when the completeness condition is violated.

When $p = 3$, without loss of generality, we assume $h_1 > 0$, $h_2 > 0$, and $h_3 < 0$ satisfying the condition $\sum_{y=1}^{3} h_y p(y \mid \delta = 1, z) = 0$ for all $z \in \{1, \dots, q\}$. Employing Lemma A.1 with $a = h_1$, $b = h_2$, $c = -h_3$, and $r_1 = r_2 = 1$, we derive:

$$\frac{1}{\pi_1^{(1)}} - \frac{1}{\pi_1^{(2)}} = h_1, \quad \frac{1}{\pi_2^{(1)}} - \frac{1}{\pi_2^{(2)}} = h_2, \quad \frac{1}{\pi_3^{(1)}} - \frac{1}{\pi_3^{(2)}} = h_3,$$

where $\sum_{j=1}^3 \pi_j^{(1)} = \sum_{j=1}^3 \pi_j^{(2)} = 1$. Substituting h_1, h_2, and h_3 into $\sum_{y=1}^3 h_y p(y \mid \delta = 1, z) = 0$ shows that the model is unidentifiable.

Lastly, we consider the case of $p \geq 4$. Suppose $h_y\,(y = 1, \dots, p)$ satisfies (39). Within $(h_1, \cdots, h_p)$, we select three elements with signs as positive, positive, and negative, respectively, and define them as a, b, and $-c$ where $a, b, c > 0$, and λ is set to be sufficiently large to ensure that

$$\lambda > 2 \max \left\{ \frac{a+b}{ab}, \frac{1}{c} \right\}. \tag{40}$$

For ease of notation, we denote $(h_1, \cdots, h_p) = (h_1, \cdots, h_{p-3}, a, b, -c)$. The remaining part of the proof is similar when the combination of the signs is negative, negative, and positive. With the selected λ, $0 < \pi_y^{(k)} < 1\,(y = 1, \dots, p-3; k = 1, 2)$ are determined to be sufficiently small to satisfy

$$\left(1 - \sum_{y=1}^{p-3} \pi_y^{(1)}\right)\left(1 - \sum_{y=1}^{p-3} \pi_y^{(2)}\right) \geq \frac{1}{2}, \quad \sum_{y=1}^{p-3} \pi_y^{(1)} < 1, \quad \sum_{y=1}^{p-3} \pi_y^{(2)} < 1, \tag{41}$$
$$\frac{1}{\pi_y^{(1)}} - \frac{1}{\pi_y^{(2)}} = \lambda h_y, \quad \text{for } y = 1, \dots, p-3.$$

Furthermore, we define r_1 and r_2 as

$$r_1^2 = 1 - \sum_{y=1}^{p-3} \pi_y^{(1)}, \; r_2^2 = 1 - \sum_{y=1}^{p-3} \pi_y^{(2)}. \tag{42}$$

By determining the variables through these steps, it follows from (40), (41), and (42) that condition (31) with $a = \lambda a$, $b = \lambda b$, and $c = \lambda c$ is fulfilled:

$$\frac{r_1^2 - r_2^2}{r_1^2 r_2^2} \leq 2(r_1^2 - r_2^2) \leq \frac{2}{c} c < (\lambda c),$$
$$-\frac{(\lambda a)(\lambda b)}{(\lambda a) + (\lambda b)} < -\frac{ab}{a+b}\frac{2(a+b)}{ab} = -2r_1^2 r_2^2 \frac{1}{r_1^2 r_2^2} \leq -\frac{1}{r_1^2 r_2^2} < \frac{r_1^2 - r_2^2}{r_1^2 r_2^2}.$$

Therefore, by applying Lemma A.1, we demonstrate that there exist $\pi_{p-2}^{(k)}$, $\pi_{p-1}^{(k)}$, and $\pi_p^{(k)}\,(k = 1, 2)$ such that

$$\sum_{y=p-2}^{p} \pi_y^{(1)} = r_1^2, \; \sum_{y=p-2}^{p} \pi_y^{(2)} = r_2^2,$$
$$\frac{1}{\pi_{p-2}^{(1)}} - \frac{1}{\pi_{p-2}^{(2)}} = \lambda a, \quad \frac{1}{\pi_{p-1}^{(1)}} - \frac{1}{\pi_{p-1}^{(2)}} = \lambda b, \quad \frac{1}{\pi_p^{(1)}} - \frac{1}{\pi_p^{(2)}} = -\lambda c.$$

The condition (39) suggests that the constructed $\pi_y^{(k)}\,(y = 1, \dots, p; k = 1, 2)$ satisfy $\sum_{y=1}^p \pi_y^{(k)} = 1$ for $k = 1, 2$ and, for any $z \in \{1, \dots, q\}$,

$$\sum_{y=1}^{p} \left(\frac{1}{\pi_y^{(1)}} - \frac{1}{\pi_y^{(2)}}\right) p(y \mid \delta = 1, z) = \lambda \sum_{y=1}^{p} h_y p(y \mid \delta = 1, z) = 0.$$

Therefore, the model is unidentifiable.

□

Proof of Theorem 2.8. We consider when y is continuous because when y is discrete, we just need to change the integral to summation. To simplify the discussion, we consider the case where $\mathcal{S}_y = \mathbb{R}$. Let $\boldsymbol{u}$ be a fixed value. Because h and g are injective functions, it is sufficient to prove the case where $\alpha := h(\boldsymbol{u}; \boldsymbol{\alpha})$ and $\beta := g(\boldsymbol{u}; \boldsymbol{\beta})$. Therefore, our goal is to prove

$$\frac{p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma})}{\int p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma})\Psi\{\alpha + \beta m(y)\}^{-1} dy} = \frac{p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma}')}{\int p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma}')\Psi\{\alpha' + \beta' m(y)\}^{-1} dy},$$

implies $\alpha = \alpha'$, $\beta = \beta'$ and $\boldsymbol{\gamma} = \boldsymbol{\gamma}'$. Integrating both sides of the above equation with respect to y yields the equality of the denominator. Thus, we have $p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma}) = p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma}')$; this implies $\boldsymbol{\gamma} = \boldsymbol{\gamma}'$ by (C4).

Next, we consider the identification of β. Taking $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ such that they satisfy (C5), we show that

$$\int \frac{p(y \mid \boldsymbol{u}, \boldsymbol{z}_1, \delta = 1; \boldsymbol{\gamma})}{\Psi\{\alpha + \beta m(y)\}} dy = \int \frac{p(y \mid \boldsymbol{u}, \boldsymbol{z}_1, \delta = 1; \boldsymbol{\gamma})}{\Psi\{\alpha' + \beta' m(y)\}} dy, \tag{43}$$

$$\int \frac{p(y \mid \boldsymbol{u}, \boldsymbol{z}_2, \delta = 1; \boldsymbol{\gamma})}{\Psi\{\alpha + \beta m(y)\}} dy = \int \frac{p(y \mid \boldsymbol{u}, \boldsymbol{z}_2, \delta = 1; \boldsymbol{\gamma})}{\Psi\{\alpha' + \beta' m(y)\}} dy, \tag{44}$$

implies $\beta = \beta'$. It follows from (43) and (44) that

$$\begin{aligned} &\int K(y; \alpha, \alpha', \beta, \beta') p(y \mid \boldsymbol{u}, \boldsymbol{z}_1, \delta = 1; \boldsymbol{\gamma}) dy \\ &= \int K(y; \alpha, \alpha', \beta, \beta') p(y \mid \boldsymbol{u}, \boldsymbol{z}_2, \delta = 1; \boldsymbol{\gamma}) dy = 0, \end{aligned} \tag{45}$$

where $K(y; \alpha, \alpha', \beta, \beta') = \Psi^{-1}\{\alpha + \beta m(y)\} - \Psi^{-1}\{\alpha' + \beta' m(y)\}$. It remains to show that (45) implies $\beta = \beta'$ in the following two steps:

Step I. We prove that the function $K(y; \alpha, \alpha', \beta, \beta')$ has a single change of sign when $\beta \neq \beta'$. Assume that $\beta \neq \beta'$. The equation $K(y; \alpha, \alpha', \beta, \beta') = 0$ has only one solution $y^* \in \mathcal{S}_y$ satisfying $m(y^*) = (\alpha - \alpha')/(\beta' - \beta)$ because of the injectivity of the function $m(\cdot)$ and $\Psi(\cdot)$. This implies $K(y)$ has a single change of sign.

Step II. We prove that the equation (45) does not hold when $\beta = \beta'$. Without loss of generality, by Step I, we consider a case where $K(y; \alpha, \alpha', \beta, \beta') < 0$ $(y < y^*)$ and $K(y; \alpha, \alpha', \beta, \beta') > 0$ $(y > y^*)$, and $p(y \mid \boldsymbol{u}, \boldsymbol{z}_2, \delta = 1)/p(y \mid \boldsymbol{u}, \boldsymbol{z}_1, \delta = 1)$ is monotone increasing. Let c be the upper bound of the density ratio

$$c := \sup_{y < y^*} \frac{p(y \mid \boldsymbol{u}, \boldsymbol{z}_2, \delta = 1)}{p(y \mid \boldsymbol{u}, \boldsymbol{z}_1, \delta = 1)}.$$

By a property on $K(y;\alpha,\alpha',\beta,\beta')$ shown in (45), we have

$$\begin{aligned}
0 &= \int K(y;\alpha,\alpha',\beta,\beta')p(y \mid \boldsymbol{u},\boldsymbol{z}_2,\delta=1)dy \\
&= \int_{-\infty}^{y^*} K(y;\alpha,\alpha',\beta,\beta')\frac{p(y \mid \boldsymbol{u},\boldsymbol{z}_2,\delta=1)}{p(y \mid \boldsymbol{u},\boldsymbol{z}_1,\delta=1)}p(y \mid \boldsymbol{u},\boldsymbol{z}_1,\delta=1)dy \\
&\quad + \int_{y^*}^{\infty} K(y;\alpha,\alpha',\beta,\beta')\frac{p(y \mid \boldsymbol{u},\boldsymbol{z}_2,\delta=1)}{p(y \mid \boldsymbol{u},\boldsymbol{z}_1,\delta=1)}p(y \mid \boldsymbol{u},\boldsymbol{z}_1,\delta=1)dy \\
&\geq \int_{-\infty}^{y^*} cK(y;\alpha,\alpha',\beta,\beta')p(y \mid \boldsymbol{u},\boldsymbol{z}_1,\delta=1)dy + \int_{y^*}^{\infty} cK(y;\alpha,\alpha',\beta,\beta')p(y \mid \boldsymbol{u},\boldsymbol{z}_1,\delta=1)dy \\
&= c\int K(y;\alpha,\alpha',\beta,\beta')p(y \mid \boldsymbol{u},\boldsymbol{z}_1,\delta=1)dy = 0,
\end{aligned}$$

where the inequality follows from the definition of c. This results in the density ratio $p(y \mid \boldsymbol{u},\boldsymbol{z}_2,\delta=1)/p(y \mid \boldsymbol{u},\boldsymbol{z}_1,\delta=1)$ being a constant on $\mathcal{S}_y$, hence, $p(y \mid \boldsymbol{u},\boldsymbol{z}_2,\delta=1) = p(y \mid \boldsymbol{u},\boldsymbol{z}_1,\delta=1)$ on $\mathcal{S}_y$. This contradicts with (C5), thus $\beta=\beta'$.

Finally, from the strict monotonicity of Ψ, it follows that the integration

$$\int \frac{p(y \mid \boldsymbol{u},\boldsymbol{z}_1,\delta=1;\boldsymbol{\gamma})}{\Psi\{\alpha+\beta m(y)\}}dy,$$

is injective with respect to α. Therefore, equation (43) implies that $\alpha=\alpha'$.

□

Proof of Proposition 2.11. It follows from the assumption (15) that there exist $M, C > 0$ such that

$$\begin{aligned}
&\int \frac{p(y \mid \boldsymbol{x},\delta=1;\boldsymbol{\gamma})}{\Psi\{h(\boldsymbol{u};\boldsymbol{\alpha})+g(\boldsymbol{u};\boldsymbol{\beta})m(y)\}}dy \\
&\propto \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\frac{(y-h(\boldsymbol{u};\boldsymbol{\alpha})-\beta\mu(\boldsymbol{x},\boldsymbol{\kappa}))^2}{\beta^2\sigma^2}\right\}\frac{1}{\Psi(y)\exp(|y|^s)}\exp(|y|^s)dy \\
&\leq \int_{-\infty}^{-M} \exp\left\{-\frac{1}{2}\frac{(y-h(\boldsymbol{u};\boldsymbol{\alpha})-\beta\mu(\boldsymbol{x},\boldsymbol{\kappa}))^2}{\beta^2\sigma^2}\right\}C\exp(|y|^s)dy + C < \infty,
\end{aligned}$$

where $0 < s < 2$. The first and the second terms of the last equation hold by the condition (15) and the increasing assumption of Ψ, respectively. □

Next, we first prove Theorem 2.18, which is the most general case. Using Theorem 2.18, we can prove Theorem 2.12 by considering the case for which $K=1$. Theorems 2.12 and 2.18 prove corollaries 2.13 and 2.19, respectively.

Proof of Theorem 2.18. Using Bayes' theorem, we obtain

$$\begin{aligned}
&p(y \mid \boldsymbol{x};\boldsymbol{\alpha},\beta,\boldsymbol{\gamma})P(\delta=1 \mid \boldsymbol{x},y;\boldsymbol{\alpha},\beta) \\
&= \frac{p(y \mid \boldsymbol{x},\delta=1;\boldsymbol{\gamma})}{\int p(y \mid \boldsymbol{x},\delta=1;\boldsymbol{\gamma})\left\{P(\delta=1 \mid \boldsymbol{x},y;\boldsymbol{\alpha},\beta)\right\}^{-1}dy}.
\end{aligned} \tag{46}$$

When $(\boldsymbol{\alpha},\beta,\boldsymbol{\gamma})$ and $(\boldsymbol{\alpha}',\beta',\boldsymbol{\gamma}')$ yield the same observed likelihood, by integrating out y from both sides, we obtain

$$\int \frac{p(y \mid \boldsymbol{x},\delta=1;\boldsymbol{\gamma})}{P(\delta=1 \mid \boldsymbol{x},y;\boldsymbol{\alpha},\beta)}dy = \int \frac{p(y \mid \boldsymbol{x},\delta=1;\boldsymbol{\gamma}')}{P(\delta=1 \mid \boldsymbol{x},y;\boldsymbol{\alpha}',\beta')}dy.$$

Then, we obtain $p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma}) = p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma}')$ because both denominators in (46) are identical. The identification of $[y \mid \boldsymbol{x}, \delta = 1]$ reduces our identification problem as

$$\int \frac{p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma})}{P(\delta = 1 \mid \boldsymbol{x}, y; \boldsymbol{\alpha}, \beta)} dy = \int \frac{p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma})}{P(\delta = 1 \mid \boldsymbol{x}, y; \boldsymbol{\alpha}', \beta')} dy \Rightarrow (\boldsymbol{\alpha}, \beta) = (\boldsymbol{\alpha}', \beta').$$

Next, we show that $\beta = \beta'$ is sufficient to show $(\boldsymbol{\alpha}, \beta) = (\boldsymbol{\alpha}', \beta')$. Let us introduce a function

$$l(s) = \int p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma}) \frac{1}{F(s + \beta y)} dy,$$

where F denotes a logistic distribution. Here, $l(s)$ inherits strict monotonicity from $F(\cdot)$. When $\beta = \beta'$, we obtain $\boldsymbol{\alpha} = \boldsymbol{\alpha}'$ using the following relationship:

$$l\left(h(\boldsymbol{x}; \boldsymbol{\alpha})\right) = l\left(h(\boldsymbol{x}; \boldsymbol{\alpha}')\right) \;\Rightarrow\; h(\boldsymbol{x}; \boldsymbol{\alpha}) = h(\boldsymbol{x}; \boldsymbol{\alpha}') \;\Rightarrow\; \boldsymbol{\alpha} = \boldsymbol{\alpha}'.$$

When $p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma})$ belongs to a mixture of the exponential family and $P(\delta = 1 \mid \boldsymbol{x}, y; \boldsymbol{\alpha}, \beta)$ is the logistic response mechanism, it can be analogously computed

$$\begin{aligned}
&\int p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma}) \left\{P(\delta = 1 \mid \boldsymbol{x}, y; \boldsymbol{\alpha}, \beta)\right\}^{-1} \, dy \\
&= \int p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma}) \left\{1 + \exp\left(-h(\boldsymbol{x}; \boldsymbol{\alpha}) - \beta y\right)\right\} \, dy \\
&= 1 + \exp\left(-h(\boldsymbol{x}; \boldsymbol{\alpha})\right) \int \sum_{k=1}^{K} \pi_k \exp\left\{\tau_k \left(y\theta_k - b(\theta_k)\right) + c(y; \tau_k)\right\} \cdot \exp\left(-\beta y\right) \, dy \\
&= 1 + \exp\left(-h(\boldsymbol{x}; \boldsymbol{\alpha})\right) \sum_{k=1}^{K} \pi_k \exp\left(-\tau_k b(\theta_k)\right) \cdot \exp\left\{\tau_k b\left(\theta_k - \frac{\beta}{\tau_k}\right)\right\}.
\end{aligned}$$

The identification problem results in

$$\begin{aligned}
\int \frac{p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma})}{P(\delta = 1 \mid \boldsymbol{x}, y; \boldsymbol{\alpha}, \beta)} dy &= \exp\left\{-g(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma})\right\} \\
&= \exp\left\{-g(\boldsymbol{\alpha}', \beta', \boldsymbol{\gamma})\right\} = \int \frac{p(y \mid \boldsymbol{x}, \delta = 1; \boldsymbol{\gamma})}{P(\delta = 1 \mid \boldsymbol{x}, y; \boldsymbol{\alpha}', \beta')} dy,
\end{aligned}$$

where $g(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma})$ is as defined in Theorem 2.18. Because the assumption $g(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}) = g(\boldsymbol{\alpha}', \beta', \boldsymbol{\gamma}) \Rightarrow \beta = \beta'$ guarantees the identification of β, we obtain the desired identification for $(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma})$. Regarding the necessity, the identifiability of $(\boldsymbol{\alpha}, \beta)$ can clearly claim the identifiability of β. Thus, the theorem is proven. □

Proof of Theorem 2.12. By using Theorem 2.18, rearranging the equation $g(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}) = g(\boldsymbol{\alpha}', \beta', \boldsymbol{\gamma})$ provides

$$\begin{aligned}
h(\boldsymbol{x}; \boldsymbol{\alpha}) - \left\{-\tau_b(\theta) + \tau b\left(\theta - \frac{\beta}{\tau}\right)\right\} &= h(\boldsymbol{x}; \boldsymbol{\alpha}') - \left\{-\tau_b(\theta) + \tau b\left(\theta - \frac{\beta'}{\tau}\right)\right\} \\
h(\boldsymbol{x}; \boldsymbol{\alpha}) - \tau b\left(\theta - \frac{\beta}{\tau}\right) &= h(\boldsymbol{x}; \boldsymbol{\alpha}') - \tau b\left(\theta - \frac{\beta'}{\tau}\right).
\end{aligned}$$

Hence, the function above is consistent with (19). Necessity and sufficiency follow from an argument analogous to the proof of Theorem 2.18. □

Proof of Theorem 2.20. Using the function $g(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma})$ in Corollary 2.19, we consider two functions $\exp\{-g\}$, which are equal but have different parameters, as follows

$$\sum_{i=1}^{K} \pi_i \exp\left\{-h(\boldsymbol{x}; \boldsymbol{\alpha}) - \beta\mu_i(\boldsymbol{x}; \boldsymbol{\kappa}_i) + \frac{1}{2}\beta^2\sigma_i^2\right\}$$
$$= \sum_{i=1}^{K} \pi_i \exp\left\{-h(\boldsymbol{x}; \boldsymbol{\alpha}') - \beta'\mu_i(\boldsymbol{x}; \boldsymbol{\kappa}_i) + \frac{1}{2}\beta'^2\sigma_i^2\right\}.$$

It suffices to show $\beta = \beta'$ to prove the identifiability according to Corollary 2.19. By employing condition (C8) and Lemma A.2, there exists $K \times K$ permutation matrix P such that

$$P\beta\mu^{\mathcal{M}}(\boldsymbol{x}) = \beta'\mu^{\mathcal{M}}(\boldsymbol{x}), \tag{47}$$

where $\mu^{\mathcal{M}}(\boldsymbol{x}) = (\mu_1^{\mathcal{M}}(\boldsymbol{x}; \boldsymbol{\kappa}_1), \ldots, \mu_K^{\mathcal{M}}(\boldsymbol{x}; \boldsymbol{\kappa}_K))^\top$. The equation (47) leads to

$$P^n\mu^{\mathcal{M}}(\boldsymbol{x}) = P^{n-1} \cdot P\mu^{\mathcal{M}}(\boldsymbol{x})$$
$$= P^{n-1} \cdot \frac{\beta'}{\beta}\mu^{\mathcal{M}}(\boldsymbol{x}) = \cdots = \left(\frac{\beta'}{\beta}\right)^n \mu^{\mathcal{M}}(\boldsymbol{x}).$$

Note that since P is the permutation matrix, there exists $n \in \mathbb{N}$ such that $P^n = I$. Thus, there exists $n \in \mathbb{N}$ such that $(\beta'/\beta)^n = 1$, which implies that $\beta = \beta'$ or $\beta = -\beta'$. The Condition (C9) indicates $\beta = \beta'$. When $\beta = -\beta'$, equation (47) becomes $P\mu^{\mathcal{M}}(\boldsymbol{x}) = -\mu^{\mathcal{M}}(\boldsymbol{x})$, indicating that (C10) is not satisfied. Therefore, this model is identifiable when (C10) holds. □

Proof of Theorem 2.23. Following the same approach as in the proof of Theorem 2.20, we consider the following equation

$$\sum_{i=1}^{K} \pi_i \exp\left\{\left(-\alpha_0 - \beta\kappa_{0i} + \frac{1}{2}\beta^2\sigma_i^2\right) - (\alpha_1 + \beta\kappa_{1i})\,x\right\}$$
$$= \sum_{i=1}^{K} \pi_i \exp\left\{\left(-\alpha_0' - \beta'\kappa_{0i} + \frac{1}{2}\beta'^2\sigma_i^2\right) - (\alpha_1' + \beta'\kappa_{1i})\,x\right\}.$$

It suffices to show $\beta = \beta'$ to prove the identifiability according to Theorem 2.18. Using Lemma A.2, there exists a permutation matrix P such that

$$P\left(\alpha_1\mathbf{1}_K + \beta\tilde{\boldsymbol{\kappa}}\right) = \alpha_1'\mathbf{1}_K + \beta'\tilde{\boldsymbol{\kappa}},$$

where $\tilde{\boldsymbol{\kappa}} = (\kappa_{11}, \ldots, \kappa_{1K})^\top$. Note that $\mathbf{1}_K$ is an eigenvector of a permutation matrix with an eigenvalue of 1. Thus, we obtain the following equation

$$P\tilde{\boldsymbol{\kappa}} = \frac{(\alpha_1' - \alpha_1)}{\beta}\mathbf{1}_K + \frac{\beta'}{\beta}\tilde{\boldsymbol{\kappa}}. \tag{48}$$

By applying equation (48) once,

$$P^2\tilde{\boldsymbol{\kappa}} = P\left\{\frac{(\alpha_1' - \alpha_1)}{\beta}\mathbf{1}_K + \frac{\beta'}{\beta}\tilde{\boldsymbol{\kappa}}\right\} = \left(1 + \frac{\beta'}{\beta}\right)\frac{(\alpha_1' - \alpha_1)}{\beta}\mathbf{1}_K + \left(\frac{\beta'}{\beta}\right)^2\tilde{\boldsymbol{\kappa}}$$

holds, and through repeating this process n times, we get

$$P^n\tilde{\boldsymbol{\kappa}} = \left\{1 + \frac{\beta'}{\beta} + \cdots + \left(\frac{\beta'}{\beta}\right)^{n-1}\right\} \frac{\alpha_1' - \alpha_1}{\beta}\mathbf{1}_K + \left(\frac{\beta'}{\beta}\right)^n \tilde{\boldsymbol{\kappa}}.$$

Note that there exists $n \in \mathbb{N}$ such that $P^n = I$ because P is the permutation matrix. Thus, the equation $P^n\tilde{\boldsymbol{\kappa}} = \tilde{\boldsymbol{\kappa}}$ holds for some $n \in \mathbb{N}$ and the following equation is obtained

$$\left\{1 - \left(\frac{\beta'}{\beta}\right)^n\right\} \tilde{\boldsymbol{\kappa}} = \left\{1 + \frac{\beta'}{\beta} + \cdots + \left(\frac{\beta'}{\beta}\right)^{n-1}\right\} \frac{\alpha_1' - \alpha_1}{\beta}\mathbf{1}_K.$$

If $(\beta'/\beta)^n \neq 1$ holds, we have $\tilde{\boldsymbol{\kappa}} = C\mathbf{1}_K$, where C is a constant. However, this result is inconsistent with $\kappa_{1i} \neq \kappa_{1j}$ $(i \neq j)$. Hence, we obtain $(\beta'/\beta)^n = 1$, meaning that $\beta = \beta'$ or $\beta = -\beta'$. Under the first condition of Theorem 2.23, $\beta = \beta'$ is immediately apparent.

Next, we show that $\beta = \beta'$ under the second condition of Theorem 2.23. Based on the above argument, if we assume $\beta = -\beta'$, equation (48) provides

$$P\tilde{\boldsymbol{\kappa}} = \frac{(\alpha_1' - \alpha_1)}{\beta}\mathbf{1}_K - \tilde{\boldsymbol{\kappa}},$$

which contradicts the second condition of Theorem 2.23. Therefore, we obtain $\beta = \beta'$. □

Proof of Theorem 2.24. Following the same approach as in the proof of Theorem 2.20, we consider the following equation

$$\sum_{i=1}^{2} \pi_i \exp\left\{\left(-\alpha_0 - \beta\kappa_{0i} + \frac{1}{2}\beta^2\sigma_i^2\right) - (\alpha_1 + \beta\kappa_{1i})\, x\right\};$$
$$= \sum_{i=1}^{2} \pi_i \exp\left\{\left(-\alpha_0' - \beta'\kappa_{0i} + \frac{1}{2}\beta'^2\sigma_i^2\right) - (\alpha_1' + \beta'\kappa_{1i})\, x\right\}.$$

It suffices to show $\beta = \beta'$ to prove the identifiability according to Theorem 2.18. Using Lemma A.2, one of the following equations holds:

$$\begin{aligned} &\text{Case 1}: \quad \alpha_1 + \beta\kappa_{11} = \alpha_1' + \beta'\kappa_{11}, \quad \alpha_1 + \beta\kappa_{12} = \alpha_1' + \beta'\kappa_{12}; \\ &\text{Case 2}: \quad \alpha_1 + \beta\kappa_{11} = \alpha_1' + \beta'\kappa_{12}, \quad \alpha_1 + \beta\kappa_{12} = \alpha_1' + \beta'\kappa_{11}. \end{aligned}$$

Under Case 1, subtracting both equations gives $\beta(\kappa_{11}-\kappa_{12}) = \beta'(\kappa_{11}-\kappa_{12})$. Therefore, we obtain $\beta = \beta'$ from the assumption $\kappa_{11} \neq \kappa_{12}$.

Under Case 2, a similar calculation of Case 1 yields $\beta = -\beta'$. Next, we compare the constant part. If the two models are not identifiable, following two equations hold

$$-\alpha_0 - \beta\kappa_{01} + \frac{1}{2}\beta^2\sigma_1^2 + \log\pi_1 = -\alpha_0' - \beta'\kappa_{02} + \frac{1}{2}\beta'^2\sigma_2^2 + \log\pi_2,$$
$$-\alpha_0 - \beta\kappa_{02} + \frac{1}{2}\beta^2\sigma_2^2 + \log\pi_2 = -\alpha_0' - \beta'\kappa_{01} + \frac{1}{2}\beta'^2\sigma_1^2 + \log\pi_1.$$

Because $\beta = -\beta'$, rearranging the equation above leads to

$$\beta^2(\sigma_1^2 - \sigma_2^2) = 2\left(\log\pi_2 - \log\pi_1\right).$$

The above equation is inconsistent with conditions 2 and 3 of Theorem 2.24. □

The following lemma shows the linear independence of exponentials of multivariate polynomials. This result plays an important role in deriving the identification conditions for normal mixtures. A related proof exists on the Stack Exchange website, and we provide it here in a more extended form.

Lemma A.2. *Let $\boldsymbol{x}$ be the p-dimensional vector $(x_1, \ldots, x_p)^\top$, $P_1(\boldsymbol{x}), \ldots, P_n(\boldsymbol{x})$ be distinct multivariate polynomials without constant term, $R_1(\boldsymbol{x}), \ldots, R_n(\boldsymbol{x})$ be rational functions of multivariate polynomials, and a domain of all these functions be the subset of the Euclidean space $\mathbb{R}^p$ which contains an interior point. Then, the following result holds:*

$$R_1(\boldsymbol{x})e^{P_1(\boldsymbol{x})} + \cdots + R_n(\boldsymbol{x})e^{P_n(\boldsymbol{x})} = 0 \;\Rightarrow\; R_1(\boldsymbol{x}) = \cdots = R_n(\boldsymbol{x}) = 0.$$

Note that Lemma A.2 implies the linear independence of exponentials of multivariate polynomials when the functions $R_1(\boldsymbol{x}), \ldots, R_n(\boldsymbol{x})$ are constant.

Proof of Lemma A.2. We prove this through mathematical induction. The case for $n = 1$ is straightforward. Now, we assume that the case for $n = k - 1$ is true. Let

$$R_1(\boldsymbol{x})e^{P_1(\boldsymbol{x})} + \cdots + R_k(\boldsymbol{x})e^{P_k(\boldsymbol{x})} = 0,$$

where $P_1(\boldsymbol{x}), \ldots, P_k(\boldsymbol{x})$ are distinct polynomials without a constant term and the functions $R_1(\boldsymbol{x}), \ldots, R_k(\boldsymbol{x})$ are rational functions. If $R_k(\boldsymbol{x}) \neq 0$, it follows

$$\frac{R_1(\boldsymbol{x})}{R_k(\boldsymbol{x})}e^{P_1(\boldsymbol{x})-P_k(\boldsymbol{x})} + \cdots + \frac{R_{k-1}(\boldsymbol{x})}{R_k(\boldsymbol{x})}e^{P_{k-1}(\boldsymbol{x})-P_k(\boldsymbol{x})} + 1 = 0. \tag{49}$$

Since both sides are differentiable at the interior point of the domain of all functions, differentiating of equation (49) with respect to x_l gives

$$\sum_{i=1}^{k-1} \left[\frac{d}{dx_l}\left(\frac{R_i(\boldsymbol{x})}{R_k(\boldsymbol{x})}\right) + \frac{R_i(\boldsymbol{x})}{R_k(\boldsymbol{x})} \cdot \frac{d}{dx_l}\{P_i(\boldsymbol{x}) - P_k(\boldsymbol{x})\}\right] e^{P_i(\boldsymbol{x})-P_k(\boldsymbol{x})} = 0.$$

Because $P_1 - P_k, \ldots, P_{k-1} - P_k$ are distinct multivariate polynomials without a constant term, the assumption of $n = k - 1$ case yields

$$\frac{d}{dx_l}\left(\frac{R_i(\boldsymbol{x})}{R_k(\boldsymbol{x})}\right) + \frac{R_i(\boldsymbol{x})}{R_k(\boldsymbol{x})} \cdot \frac{d}{dx_l}(P_i(\boldsymbol{x}) - P_k(\boldsymbol{x})) = 0,$$
$$\left\{\frac{d}{dx_l}\left(\frac{R_i(\boldsymbol{x})}{R_k(\boldsymbol{x})}\right) + \frac{R_i(\boldsymbol{x})}{R_k(\boldsymbol{x})} \cdot \frac{d}{dx_l}(P_i(\boldsymbol{x}) - P_k(\boldsymbol{x}))\right\} e^{P_i(\boldsymbol{x})-P_k(\boldsymbol{x})} = 0,$$

where $i = 1, 2, \ldots, k - 1$. Integrating out x_l from both sides, we obtain

$$\frac{R_i(\boldsymbol{x})}{R_k(\boldsymbol{x})}e^{P_i(\boldsymbol{x})-P_k(\boldsymbol{x})} = C_i(x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_p),$$

where $l = 1, 2, \cdots, p$. The left-hand side is constant because it does not depend on l. Therefore, it is denoted by C_i. If $C_i \neq 0$, $R_i(\boldsymbol{x})/R_k(\boldsymbol{x})$ and $P_i(\boldsymbol{x}) - P_k(\boldsymbol{x})$ are constants, which contradicts the fact that $P_1(\boldsymbol{x}), \ldots, P_n(\boldsymbol{x})$ are distinct polynomials without constant terms. Thus, using $C_i = 0$ for $i = 1, 2, \ldots, k - 1$ and the equation (49), the following contradictory equation holds

$$0 = \frac{R_1(\boldsymbol{x})}{R_k(\boldsymbol{x})}e^{P_1(\boldsymbol{x})-P_k(\boldsymbol{x})} + \cdots + \frac{R_{k-1}(\boldsymbol{x})}{R_k(\boldsymbol{x})}e^{P_{k-1}(\boldsymbol{x})-P_k(\boldsymbol{x})} + 1 = 1.$$

Consequently, we obtain $R_k(\boldsymbol{x}) = 0$ that follows

$$R_1(\boldsymbol{x})e^{P_1(\boldsymbol{x})} + \cdots + R_{k-1}(\boldsymbol{x})e^{P_{k-1}(\boldsymbol{x})} = 0.$$

From the assumption of $n = k - 1$, $R_1(\boldsymbol{x}) = \cdots = R_{k-1}(\boldsymbol{x}) = 0$ holds and the lemma is proven. $\square$

B Technical Proofs of Section 3

The following are additional conditions to ensure consistency and asymptotic normality.

Condition 7. *Let $\alpha_j\,(j=1,\ldots,J)$, $\beta_k\,(k=1,\ldots,K)$ and $\gamma_l\,(l=1,\ldots,L)$ be the parameter of candidate working models $\pi^{[j]}$, $g^{[k]}$, and $C^{[l]}(X)$. We use $\hat\alpha_k$, $\hat\beta_j$, and $\hat\gamma_l$ to denote the estimators of α_k, β_j, and γ_l. $N^{1/2}(\hat\alpha_k-\alpha_k^*)$, $N^{1/2}(\hat\beta_j-\beta_j^*)$, and $N^{1/2}(\hat\gamma_l-\gamma_l^*)$ are bounded in probability, where α_k^*, β_j^*, and γ_l^* are the probability limit of $\hat\alpha_k$, $\hat\beta_j$, and $\hat\gamma_l$. Let $\alpha^*=(\alpha_1^{*\top},\ldots,\alpha_J^{*\top})^\top$, $\beta^*=(\beta_1^{*\top},\ldots,\beta_K^{*\top})^\top$, and $\gamma^*=(\gamma_1^{*\top},\ldots,\gamma_L^{*\top})^\top$.*

Condition 8. *$E(W^3)<\infty$.*

Condition 9. *Suppose that $\pi^{[1]}(x,z,w;\alpha_1)$ and $g^{[1]}(x,z,w;\beta_1)$ are correctly specified. Then, functions*

$$
\begin{aligned}
&\|R\delta W(1-\delta W)U_\theta/\pi^{[1]}\|,\ \|\partial^2\{\hat h^\top(\alpha,\beta)/\pi^{[1]}\}/\partial(\alpha^\top,\beta^\top)\partial(\alpha^\top,\beta^\top)^\top\|,\ \|\hat h^\top(\alpha,\beta)/\pi^{[1]}\|^3,\\
&\|\partial\{\hat h^\top(\alpha,\beta)/\pi^{[1]}(X,Z,W;\alpha_1)\}/\partial(\alpha^\top,\beta^\top)\|,\ \|(1+\hat\rho^\top\hat h)^{-1}R\delta W\left\{U_{\theta^*}(X,Y)-g^{[1]}\right\}\|,\\
&\|R\delta W U_\theta(X,Y)/\pi^{[1]}\|,\ \|(1-\delta W)^2\mathcal{C}_\theta(X;\gamma)^{\otimes 2}\|,\ \|R\hat h^{\otimes 2}/\{(\pi^{[1]})^2W\}\|,\\
&\|\delta R(\pi^{[1]})^{-1}\partial\hat h/\partial\alpha_1^\top\|,\ \|\delta R(\pi^{[1]})^{-2}\hat h\partial\pi^{[1]}/\partial\alpha_1\|,\ \|\delta R\hat h/\pi^{[1]}\|,\ \|R\delta W U_\theta(X,Y)\hat h^\top/(\pi^{[1]})^2\|,\\
&\|R\delta W(W-1)U_\theta(X,Y)\mathcal{C}_\theta(X;\gamma)^\top/\pi^{[1]}\|,\ \|R\delta W(\pi^{[1]})^{-2}U_\theta(X,Y)\partial\pi^{[1]}/\partial\alpha_1^\top\|,\\
&\|R\delta W/\pi^{[1]}\{\partial U_\theta(X,Y)/\partial\theta^\top\}\|,
\end{aligned}
$$

are continuous and bounded by some integrable function in the neighborhood of the point $(\alpha^{*\top},\beta^{*\top},\gamma^{*\top},\theta^{*\top})$, where $\hat h(\alpha,\beta,\theta)^\top=(\hat h_1^\top,\hat h_2^\top)$,

$$
\begin{aligned}
\hat h_1^\top &= (\pi^{[1]}(\alpha_1)-\bar\pi_n^{[1]}(\alpha_1),\ldots,\pi^{[J]}(\alpha_J)-\bar\pi_n^{[J]}(\alpha_J)),\\
\hat h_2^\top &= (Wg_\theta^{[1]}(\beta_1)^\top-\bar g_\theta^{w[1]}(\beta_1)^\top,\ldots,Wg^{[K]}(\beta_K)^\top-\bar g_\theta^{w[K]}(\beta_K)^\top),
\end{aligned}
$$

ρ^* is the probability limit of the Lagrange multipliers $\hat\rho$ satisfied (50), and $\mathcal{C}_\theta(X;\gamma)=(C_\theta^{[1]}(X;\gamma^{[1]})^\top,\ldots,C_\theta^{[L]}(X;\gamma^{[L]})^\top)^\top$.

Condition 10. *Functions $\|(1-\delta W)\mathcal{C}_\theta(X;\gamma)\|^3$, $\|\partial\{(1-\delta W)\mathcal{C}_\theta(X;\gamma)\}/\partial(\theta^\top,\gamma^\top)\|$, and $\|\partial^2\{(1-\delta W)\mathcal{C}_\theta(X;\gamma)\}/\partial(\theta^\top,\gamma^\top)\partial(\theta^\top,\gamma^\top)^\top\|$ are continuous and bounded by some integrable function with respect to the probability distribution (X,Z,W) in the neighborhood of $(\gamma^{*\top},\theta^{*\top})$.*

Condition 11. *Functions*

$$
\|\partial^2\{WD_\tau^*(X_i,Z_i)\}/\partial\tau^\top\partial\tau\|,\ \|\partial\{WD_\tau^*(X_i,Z_i)\}/\partial\tau^\top\|,\ \|WD_\tau^*(X_i,Z_i)\|^3,
$$

are continuous and bounded by some integrable function in the neighborhood of τ^*, and

$$
\begin{aligned}
&\|\delta RW^2U_\theta(X,Y)D_\tau^*(X,Z)^\top/\pi^{[1]}\|,\ \|R\delta W/\pi^{[1]}\{\partial U_\theta(X,Y)/\partial\theta^\top\}\|,\ \|(W-1)D_\tau^*\|,\\
&\|W(D_\tau^*)^{\otimes 2}\|,\ \|\partial D_\tau^*/\partial\tau\|,
\end{aligned}
$$

are bounded by some integrable function in the neighborhood of $(\alpha_1^{*\top},\theta^{*\top},\tau^{*\top})$.

Proof of Theorem 3.4. For each Setting 1 and 2, we show multiple robustness and then prove the semiparametric efficiency. Suppose that $\pi^{[1]}$ is the correct model.

By using the method of Lagrange multipliers, the first-step empirical weights are obtainable as

$$\hat{p}_i^{(1)} = m^{-1} \frac{1}{1 + \hat{\rho}^\top \hat{h}_i(\hat{\alpha}, \hat{\beta}, \hat{\theta}_1)}, \quad (i = 1, \ldots, m)$$

where

$$\hat{h}_i(\hat{\alpha}, \hat{\beta}, \hat{\theta}_1) = \Big(\hat{\pi}^{[1]}(X_i, Z_i, W_i; \hat{\alpha}_1) - \bar{\pi}_n^{[1]}(\hat{\alpha}_1), \ldots, \hat{\pi}^{[J]}(X_i, Z_i, W_i; \hat{\alpha}_J) - \bar{\pi}_n^{[J]}(\hat{\alpha}_J),$$
$$W_i \hat{g}_{\hat{\theta}_1}^{[1]}(X_i, Z_i, W_i; \hat{\beta}_1) - \bar{g}_{\hat{\theta}_1}^{w[1]}(\hat{\beta}_1), \ldots, W_i \hat{g}_{\hat{\theta}_1}^{[K]}(X_i, Z_i, W_i; \hat{\beta}_K) - \bar{g}_{\hat{\theta}_1}^{w[K]}(\hat{\beta}^K)\Big)^\top,$$

$\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_J)^\top$, $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_K)^\top$, and the Lagrange multipliers $\hat{\rho}$ satisfy

$$m^{-1} \sum_{i=1}^m \frac{\hat{h}_i(\hat{\alpha}, \hat{\beta}, \hat{\theta}_1)}{1 + \hat{\rho}^\top \hat{h}_i(\hat{\alpha}, \hat{\beta}, \hat{\theta}_1)} = 0. \tag{50}$$

Next, we consider empirical weights by using the information that $\pi^{[1]}$ is correct, as considered in Han (2014): $\max_{q_1,\ldots,q_m} \prod_{i=1}^m q_i$, subject to

$$\sum_{i=1}^m q_i^{(1)} = 1,$$
$$\sum_{i=1}^m q_i^{(1)} \left\{\hat{\pi}^{[j]}(X_i, Z_i, W_i) - \bar{\pi}_n^{[j]}\right\} / \hat{\pi}^{[1]}(X_i, Z_i, W_i) = 0, \quad (j = 1, \ldots, J)$$
$$\sum_{i=1}^m q_i^{(1)} \left\{W_i \hat{g}_\theta^{[k]}(X_i, Z_i, W_i) - \bar{g}_\theta^{w[k]}\right\} / \hat{\pi}^{[1]}(X_i, Z_i, W_i) = 0. \quad (k = 1, \ldots, K)$$

By using the method of Lagrange multipliers again, we obtain

$$\hat{q}_i = m^{-1} \frac{1}{1 + \hat{\lambda}^\top \hat{h}_i(\hat{\alpha}, \hat{\beta}, \hat{\theta}) / \hat{\pi}^{[1]}(X_i, Z_i, W_i)}, \quad (i = 1, \ldots, m)$$

where the Lagrange multipliers $\hat{\lambda}$ satisfy

$$m^{-1} \sum_{i=1}^m \frac{\hat{h}_i(\hat{\alpha}, \hat{\beta}, \hat{\theta}) / \hat{\pi}^{[1]}(X_i, Z_i, W_i)}{1 + \hat{\lambda}^\top \hat{h}_i(\hat{\alpha}, \hat{\beta}, \hat{\theta}) / \hat{\pi}^{[1]}(X_i, Z_i, W_i)} = 0. \tag{51}$$

It follows from the equation

$$m^{-1} \sum_{i=1}^m \frac{\hat{h}_i / \hat{\pi}_i^{[1]}}{1 + \lambda^\top \hat{h}_i / \hat{\pi}_i^{[1]}} = (\bar{\pi}_n^{[1]} m)^{-1} \sum_{i=1}^m \frac{\hat{h}_i}{1 + (\lambda_1 + 1, \lambda_2, \ldots, \lambda_{J+K}) \hat{h}_i / \bar{\pi}_n^{[1]}},$$

that the Lagrange multipliers $\hat{\rho}$ are $\hat{\rho}_1 = (\hat{\lambda}_1 + 1)/\bar{\pi}_n^{[1]}$ and $\hat{\rho}_l = \hat{\lambda}_l / \bar{\pi}_n^{[1]}$, $l = 2, \ldots, J + K$. Thus, we have

$$\hat{p}_i^{(1)} = m^{-1} \frac{\bar{\pi}_n^{[1]} / \hat{\pi}^{[1]}(X_i, Z_i, W_i)}{1 + \hat{\lambda}^\top \hat{h}_i / \hat{\pi}^{[1]}(X_i, Z_i, W_i)},$$

and each $\hat{\theta}_1$ and $\hat{\lambda}$ converges to θ^* and 0 in probability, respectively, from the standard theory of empirical-likelihood method in Qin and Lawless (1994).

In the second step, the empirical weights are written as

$$\hat{p}_i^{(2)} = N^{-1} \frac{1}{1 + \hat{\upsilon}^\top (1 - \delta_i W_i) \mathcal{C}_{\hat{\theta}_2}(X_i; \hat{\gamma})},$$

where $\mathcal{C}_{\hat{\theta}_2}(X_i;\hat{\gamma}) = (\hat{C}_{\hat{\theta}_2}^{[1]}(X_i;\hat{\gamma}^{[1]})^\top, \ldots, \hat{C}_{\hat{\theta}_2}^{[L]}(X_i;\hat{\gamma}^{[L]})^\top)^\top$, $\hat{\gamma} = (\hat{\gamma}^{[1]}, \ldots, \hat{\gamma}^{[L]})^\top$, and the Lagrange multipliers $\hat{\upsilon}$ satisfy

$$\sum_{i=1}^{N} \frac{(1 - \delta_i W_i) \mathcal{C}_{\hat{\theta}_2}(X_i;\hat{\gamma})}{1 + \hat{\upsilon}^\top (1 - \delta_i W_i) \mathcal{C}_{\hat{\theta}_2}(X_i;\hat{\gamma})} = 0. \tag{52}$$

By using the theory in Qin and Lawless (1994) again, we can show that each $\hat{\theta}_2$ and $\hat{\upsilon}$ converges to θ^* and 0 in probability. Therefore, the two empirical weights and the uniform law of large numbers provide the asymptotic unbiasedness of the estimating equation:

$$\begin{aligned}
& n \sum_{i=1}^{N} \hat{p}_i^{(2)} \hat{p}_i^{(1)} \delta_i R_i W_i U_{\hat{\theta}_{\mathrm{EL1}}}(X_i, Y_i) \\
&= n \frac{\bar{\pi}_n^{[1]}}{m} N^{-1} \sum_{i=1}^{N} \frac{1}{1 + \hat{\upsilon}^\top (1 - \delta_i W_i) \mathcal{C}_{\hat{\theta}_2}(X_i;\hat{\gamma})^\top} \frac{R_i / \hat{\pi}^{[1]}(X_i, Z_i, W_i)}{1 + \hat{\lambda}^\top \hat{h}_i / \hat{\pi}^{[1]}(X_i, Z_i, W_i)} \delta_i W_i U_{\hat{\theta}_{\mathrm{EL1}}}(X_i, Y_i) \\
&= N^{-1} \sum_{i=1}^{N} \frac{R_i}{\hat{\pi}^{[1]}(X_i, Z_i, W_i)} \delta_i W_i U_{\hat{\theta}_{\mathrm{EL1}}}(X_i, Y_i) + o_p(1) \\
&\xrightarrow{p} E\left\{ \frac{R}{\pi^{[1]}(X, Z, W)} \delta W U_{\theta^*}(X, Y) \right\} = 0.
\end{aligned}$$

Next, we consider when the outcome models include the correct model and suppose that $g^{[1]}$ is the correct model. It follows from the constraint

$$\sum_{i=1}^{N} \hat{p}_i^{(1)} R_i W_i g^{[1]}(X_i, Z_i, W_i) = n^{-1} \sum_{i=1}^{n} W_i g^{[1]}(X_i, Z_i, W_i) \tag{53}$$

that the estimating equation is asymptotically unbiased:

$$\begin{aligned}
& N \sum_{i=1}^{N} \hat{p}_i^{(2)} \hat{p}_i^{(1)} \delta_i R_i W_i U_{\hat{\theta}_{\mathrm{EL1}}}(X_i, Y_i) \\
&= \sum_{i=1}^{N} \hat{p}_i^{(1)} R_i \left\{ \delta_i W_i U_{\hat{\theta}_{\mathrm{EL1}}}(X_i, Y_i) - W_i g^{[1]}(X_i, Z_i, W_i) \right\} + n^{-1} \sum_{i=1}^{n} W_i g^{[1]}(X_i, Z_i, W_i) + o_p(1) \\
&= m^{-1} \sum_{i=1}^{N} \frac{1}{1 + \hat{\rho}^\top \hat{h}_i} R_i \delta_i W_i \left\{ U_{\hat{\theta}_{\mathrm{EL1}}}(X_i, Y_i) - g^{[1]}(X_i, Z_i, W_i) \right\} + o_p(1) \\
&\xrightarrow{p} \frac{1}{P(\delta = R = 1)} E\left[\frac{R \delta W \left\{ U_{\theta^*}(X, Y) - g^{[1]}(X, Z, W) \right\}}{1 + \rho^{*\top} h^\dagger(X, Z, W)} \right] = 0,
\end{aligned}$$

where

$$\begin{aligned}
h^\dagger(X, Z, W)^\top = (&\pi^{[1]}(\alpha_1^*) - \Gamma_1, \ldots, \pi^{[J]}(\alpha_J^*) - \Gamma_J, \\
& W g_{\theta_*}^{[1]}(X, Z, W; \beta_1^*) - \Gamma_{J+1}, \ldots, W_i g_{\theta_*}^{[K]}(X, Z, W; \beta_K^*) - \Gamma_{J+K}),
\end{aligned}$$

and $\Gamma_j\,(j=1,\ldots,J)$ and $\Gamma_{J+k}\,(k=1,\ldots,K)$ are the probability limit of $\bar{\pi}_n^{[j]}\,(j=1,\ldots,J)$ and $\bar{g}_{\hat{\theta}_1}^{w[k]}\,(k=1,\ldots,K)$. Therefore, the estimator $\hat{\theta}_{\mathrm{EL1}}$ has consistency when one of the $J+K$ models for the response mechanism and the regression function is correct.

We prove the efficiency of $\hat{\theta}_{\mathrm{EL1}}$. The Taylor expansion of the left-hand side of (52) around $(\upsilon^\top,\gamma^\top,\theta^\top)=(0^\top,\gamma^{*\top},\theta^{*\top})$ is

$$\begin{aligned}0 &= N^{-1/2}\sum_{i=1}^N (1-\delta_i W_i)\mathcal{C}_{\theta^*}(X_i;\gamma^*)\\ &\quad - N^{-1}\sum_{i=1}^N (1-\delta_i W_i)^2\mathcal{C}_{\theta^*}(X_i;\gamma^*)^{\otimes 2} N^{1/2}\hat{\upsilon} + o_p(1).\end{aligned}$$

With this equation, $\hat{\upsilon}$ can be expanded as

$$N^{1/2}\hat{\upsilon} = E\left\{(W-1)\mathcal{C}_{\theta^*}(X;\gamma^*)^{\otimes 2}\right\}^{-1} N^{-1/2}\sum_{i=1}^N (1-\delta_i W_i)\mathcal{C}_{\theta^*}(X_i;\gamma^*) + o_p(1). \tag{54}$$

In a similar way, the left-hand side of (51) can be expanded as

$$\begin{aligned}0 &= N^{-1/2}\sum_{i=1}^N \delta_i \frac{R_i-\pi_i^{[1]}}{\pi_i^{[1]}} h^\dagger(X_i,Z_i,W_i) - E\left\{\frac{h^\dagger(X,Z,W)^{\otimes 2}}{\pi^{[1]}W}\right\} N^{1/2}\hat{\lambda}\\ &\quad + N^{-1/2}\sum_{i=1}^N \delta_i R_i\left\{\frac{1}{\left(\pi_i^{[1]}\right)^2}\left(\frac{\partial \hat{h}_i}{\partial\alpha_1^\top}\pi_i^{[1]} - \hat{h}_i\frac{\partial \pi_i^{[1]}}{\partial \alpha_1}\right) - \frac{\hat{h}_i}{\pi_i^{[1]}}1^\top\right\}(\hat{\alpha}_1-\alpha^*) + o_p(1). \end{aligned}\tag{55}$$

Recall that our empirical-likelihood estimator is the solution to

$$Q(\hat{\upsilon},\hat{\lambda},\hat{\alpha},\hat{\beta},\hat{\gamma},\hat{\theta}_1,\hat{\theta}_2,\hat{\theta}_{\mathrm{EL1}}) = \sum_{i=1}^N \hat{p}_i^{(2)}(\hat{\upsilon},\hat{\gamma},\hat{\theta}_2)\hat{p}_i^{(1)}(\hat{\lambda},\hat{\alpha},\hat{\beta},\hat{\theta}_1)\delta_i R_i W_i U_{\hat{\theta}_{\mathrm{EL1}}}(X_i,Y_i) = 0.$$

By substituting $p_i^{(1)}$ and $p_i^{(2)}$ into the above equation and expanding, we have

$$\begin{aligned}0 &= N^{-1/2}\sum_{i=1}^N \left(1-\frac{\hat{\lambda}^\top\hat{h}_i/\hat{\pi}_i^{[1]}}{1+\hat{\lambda}^\top\hat{h}_i/\hat{\pi}_i^{[1]}}\right)\left\{1-\frac{\hat{\upsilon}^\top(1-\delta_i W_i)\mathcal{C}_{\hat{\theta}_2}(X_i;\hat{\gamma})}{1+\hat{\upsilon}^\top(1-\delta_i W_i)\mathcal{C}_{\hat{\theta}_2}(X_i;\hat{\gamma})}\right\}\\ &\qquad \cdot \delta_i W_i \frac{R_i}{\hat{\pi}_i^{[1]}} U_{\hat{\theta}_{\mathrm{EL1}}}(X_i,Y_i)\\ &= N^{-1/2}\sum_{i=1}^N \delta_i W_i\frac{R_i}{\hat{\pi}_i^{[1]}}U_{\hat{\theta}_{\mathrm{EL1}}}(X_i,Y_i) - N^{-1}\sum_{i=1}^N \delta_i W_i \frac{R_i}{\hat{\pi}_i^{[1]}}\frac{U_{\hat{\theta}_{\mathrm{EL1}}}(X_i,Y_i)\hat{h}_i^\top/\hat{\pi}_i^{[1]}}{1+\hat{\lambda}^\top\hat{h}_i/\hat{\pi}_i^{[1]}}N^{1/2}\hat{\lambda}\\ &\quad + N^{-1}\sum_{i=1}^N \delta_i W_i\frac{R_i}{\hat{\pi}_i^{[1]}}\frac{(W_i-1)U_{\hat{\theta}_{\mathrm{EL1}}}(X_i,Y_i)\mathcal{C}_{\hat{\theta}}^\top(X_i;\hat{\gamma})}{1+\hat{\upsilon}^\top(1-\delta_i W_i)\mathcal{C}_{\hat{\theta}}(X_i;\hat{\gamma})}N^{1/2}\hat{\upsilon} + o_p(1)\\ &= N^{-1/2}\sum_{i=1}^N \frac{R_i}{\pi_i^{[1]}}\delta_i W_i U_{\theta^*}(X_i,Y_i) - E\left\{\frac{U_{\theta^*}(X,Y)}{\pi^{[1]}}h^\dagger(X_i,Z_i,W_i)^\top\right\}N^{1/2}\hat{\lambda}\\ &\quad + E\left\{(W-1)U_{\theta^*}(X,Y)\mathcal{C}_{\theta^*}(X;\gamma^*)^\top\right\}N^{1/2}\hat{\upsilon}\\ &\quad + E\left\{\frac{U_{\theta^*}(X,Y)}{\pi^{[1]}}\left(\frac{\partial \pi^{[1]}(\alpha_1^*)}{\partial\alpha_1}\right)^\top\right\}N^{1/2}\left(\hat{\alpha}_1^*-\alpha_1^*\right)\\ &\quad + E\left\{\frac{\partial U_{\theta^*}(X,Y)}{\partial\theta}\right\}N^{1/2}\left(\hat{\theta}_{\mathrm{EL1}}-\theta^*\right) + o_p(1). \end{aligned}\tag{56}$$

It remains to show that terms in (56) reduce to

$$
\begin{aligned}
&E\left\{(W-1)U_{\theta^*}(X,Y)\mathcal{C}_{\theta^*}(X;\gamma^*)^\top\right\}N^{1/2}\hat{\upsilon}\\
&=E\left\{(W-1)C_{\theta^*}^{[1]}\mathcal{C}_{\theta^*}^\top\right\}E\left\{(W-1)\mathcal{C}_{\theta^*}^{\otimes 2}\right\}^{-1}N^{-1/2}\sum_{i=1}^N(1-\delta_iW_i)\mathcal{C}_{\theta^*}(X_i;\gamma^*)+o_p(1)\\
&=N^{-1/2}\sum_{i=1}^N(1-\delta_iW_i)C_{\theta^*}^{[1]}(X_i)+o_p(1),
\end{aligned}
\tag{57}
$$

and

$$
\begin{aligned}
&E\left\{\frac{U_{\theta^*}(X,Y)}{\pi^{[1]}}h^\dagger(X_i,Z_i,W_i)^\top\right\}N^{1/2}\hat{\lambda}\\
&=E\left(\frac{Wg^{[1]}}{\pi^{[1]}W}h^{\dagger\top}\right)E\left(\frac{h^{\dagger\otimes 2}}{\pi^{[1]}W}\right)^{-1}\left[N^{-1/2}\sum_{i=1}^N\delta_i\frac{R_i-\pi_i^{[1]}}{\pi_i^{[1]}}h_i^\dagger\right.\\
&\quad\left.+N^{-1}\sum_{i=1}^N\delta_iR_i\left\{\frac{1}{\left(\pi_i^{[1]}\right)^2}\left(\frac{\partial\hat{h}_i}{\partial\alpha_1^\top}\pi_i^{[1]}-\hat{h}_i\frac{\partial\pi_i^{[1]}}{\partial\alpha_1}\right)-\frac{\hat{h}_i}{\pi_i^{[1]}}1^\top\right\}N^{1/2}(\hat{\alpha}_1-\alpha^*)\right]+o_p(1)\\
&=(0,\ldots,0,1,0,\ldots,0)\left[N^{-1/2}\sum_{i=1}^N\delta_i\frac{R_i-\pi_i^{[1]}}{\pi_i^{[1]}}h_i^\dagger\right.\\
&\quad\left.+N^{-1}\sum_{i=1}^N\delta_iR_i\left\{\frac{1}{\left(\pi_i^{[1]}\right)^2}\left(\frac{\partial\hat{h}_i}{\partial\alpha_1^\top}\pi_i^{[1]}-\hat{h}_i\frac{\partial\pi_i^{[1]}}{\partial\alpha_1}\right)-\frac{\hat{h}_i}{\pi_i^{[1]}}1^\top\right\}N^{1/2}(\hat{\alpha}_1-\alpha^*)\right]+o_p(1)\\
&=N^{-1/2}\sum_{i=1}^N\delta_i\frac{R_i-\pi_i^{[1]}}{\pi_i^{[1]}}g_i^{[1]}+E\left\{\frac{U_{\theta^*}}{\pi^{[1]}}\left(\frac{\partial\pi^{[1]}(\alpha_1^*)}{\partial\alpha_1}\right)^\top\right\}N^{1/2}(\hat{\alpha}_1-\alpha^*)+o_p(1).
\end{aligned}
\tag{58}
$$

Then, equations (56)–(58) reveal that the influence function of the two-step empirical-likelihood estimator in Setting 1 is asymptotically the same as $S_{\text{eff},1}$.

We prove the property of $\hat{\theta}_{\text{EL2}}$. In setting 2, by using the method of Lagrange multipliers, the second-step empirical weights are

$$
\hat{p}_i^{(2)}=n^{-1}\frac{1}{1+\hat{\zeta}(W_i^{-1}-\hat{V})},\quad (i=1,\ldots,n)
$$

where $\hat{\zeta}$ and $\hat{V}$ satisfy

$$
\sum_{i=1}^n\frac{\zeta(1/W_i-V)}{1+\zeta(1/W_i-V)}=0,\quad \sum_{i=1}^n\frac{\zeta}{1+\zeta(1/W_i-V)}-\frac{N-n}{1-V}=0.
\tag{59}
$$

It follows from (59) that $\hat{V}=1+\hat{\zeta}^{-1}(1-N/n)$.

First, we prove the consistency of $\hat{\zeta}$ and $\hat{V}$. By using the same arguments of Qin, Leung and Shao (2002), after profiling out $p_i^{(2)}$, the log-likelihood is $l_1(\xi,V)+l_2(V)$, where $l_1(\zeta,V)=-\sum_{i=1}^n\log\{1+\xi V(1-VW_i)\}$, $l_2(V)=n\log VW_i+(N-n)\log(1-V)$, and $\xi=\zeta-V^{-1}$ satisfy

$$
\sum_{i=1}^n\frac{V\,(1-VW_i)}{1+\xi V\,(1-VW_i)}=0.
$$

By a similar argument to Qin et al. (2002), if V is in the set $\{V : \|V\| = N^{-1/3}\}$, we can show that $l_1(\xi(V), V) > l_1(\xi(V^*), V^*)$ a.s. and $l_1(\xi(V), V) + l_2(W) > l_1(\xi(V^*), V^*) + l_2(W_0)$ a.s. Therefore, we obtain consistency of $\hat{\zeta} \to P(\delta = 1)^{-1}$ and $\hat{V} \to P(\delta = 1)$.

It follows analogously from

$$N \sum_{i=1}^{N} \hat{p}_i^{(2)}(\hat{\zeta}) \hat{p}_i^{(1)}(\hat{\lambda}, \hat{\alpha}, \hat{\beta}, \hat{\theta}_1) \delta_i R_i U_{\hat{\theta}_{\mathrm{EL2}}}(X_i, Y_i) = \sum_{i=1}^{N} \hat{p}_i^{(1)}(\hat{\lambda}, \hat{\alpha}, \hat{\beta}, \hat{\theta}_1) \delta_i R_i W_i U_{\hat{\theta}_{\mathrm{EL2}}}(X_i, Y_i),$$

that our empirical-likelihood-based estimator in Setting 2 has multiple robustness.

By using the Taylor expansion, we obtain

$$N^{1/2}\left\{\hat{\zeta} - P(\delta = 1)^{-1}\right\} = \frac{1}{P(\delta = 1) E(W - 1)} N^{-1/2} \sum_{i=1}^{N} (1 - \delta_i W_i) + o_p(1). \tag{60}$$

Recall that our empirical-likelihood estimator is the solution to

$$Q(\hat{\zeta}, \hat{\lambda}, \hat{\alpha}, \hat{\beta}, \hat{\theta}_1, \hat{\theta}_{\mathrm{EL2}}) = \sum_{i=1}^{N} \hat{p}_i^{(2)}(\hat{\zeta}) \hat{p}_i^{(1)}(\hat{\lambda}, \hat{\alpha}, \hat{\beta}, \hat{\theta}_1) \delta_i R_i U_{\hat{\theta}_{\mathrm{EL2}}}(X_i, Y_i) = 0.$$

The Taylor expansion of the estimating equation is

$$\begin{aligned} 0 = {} & N^{-1/2} \sum_{i=1}^{N} \delta_i W_i \left\{ \frac{R_i}{\pi_i^{[1]}} U_{\theta^*}(X_i, Y_i) + \frac{\pi_i^{[1]} - R_i}{\pi_i^{[1]}} g_{\theta^*}^{[1]}(X_i, Y_i, W_i; \beta_1^*) \right\} \\ & + P(\delta = 1) E\left\{(W - 1) U_{\theta^*}(X, Y)\right\} N^{1/2} \left\{\hat{\zeta} - P(\delta = 1)^{-1}\right\} \\ & + E\left(\frac{\partial U_{\theta^*}(X, Y)}{\partial \theta}\right) N^{1/2} \left(\hat{\theta}_{\mathrm{EL2}} - \theta^*\right) + o_p(1). \end{aligned} \tag{61}$$

By substituting (60) into the second term of (61), we have

$$\begin{aligned} & - E\left(\frac{\partial S_{\mathrm{eff},2}}{\partial \theta^\top}\right) N^{1/2} \left(\hat{\theta}_{\mathrm{EL2}} - \theta^*\right) \\ & = N^{-1/2} \sum_{i=1}^{N} \delta_i W_i \left\{ \frac{R_i}{\pi_i^{[1]}} U_{\theta^*}(X_i, Y_i) + \frac{\pi_i^{[1]} - R_i}{\pi_i^{[1]}} g_{\theta^*}^{[1]}(X_i, Y_i, W_i; \beta_1^*) \right\} \\ & \quad + N^{-1/2} \sum_{i=1}^{N} (1 - \delta_i W_i) \frac{E\left\{(W - 1) U_{\theta^*}(X, Y)\right\}}{E(W - 1)} + o_p(1). \end{aligned}$$

Thus, the influence function of the two-step empirical-likelihood estimator in Setting 2 is asymptotically the same as $S_{\mathrm{eff},1}$. □

Proof of Theorem 3.5. Because the proof of multiple robustness is almost the same, we prove only the semiparametric efficiency of our proposed estimator in Setting 3. In Setting 3, the empirical weights are represented by

$$\hat{p}_i^{(2)}(\nu, \zeta, \tau) = n^{-1} \frac{1}{N/n + \nu(W_i^{-1} - 1) + \zeta D_\tau^*(X_i, Z_i)}.$$

Recall that our empirical-likelihood estimator is the solution to

$$Q(\hat{\nu}, \hat{\zeta}, \hat{\tau}, \hat{\lambda}, \hat{\alpha}, \hat{\beta}, \hat{\theta}_1, \hat{\theta}_{\mathrm{EL3}}) = \sum_{i=1}^{N} \hat{p}_i^{(2)}(\hat{\nu}, \hat{\zeta}, \hat{\tau}) \hat{p}_i^{(1)}(\hat{\lambda}, \hat{\alpha}, \hat{\beta}, \hat{\theta}_1) \delta_i R_i U_{\hat{\theta}_{\mathrm{EL3}}}(X_i, Y_i) = 0.$$

The Taylor expansion of the estimating equation is

$$\begin{aligned} 0 = {} & N^{-1/2} \sum_{i=1}^{N} \delta_i W_i \left\{ \frac{R_i}{\pi_i^{[1]}} U_{\theta^*}(X_i, Y_i) + \frac{\pi_i^{[1]} - R_i}{\pi_i^{[1]}} g_{\theta^*}^{[1]}(X_i, Y_i, W_i; \beta_1^*) \right\} \\ & + P(\delta = 1) E\left\{(W-1) U_{\theta^*}(X, Y)\right\} N^{1/2} \left\{\hat{\nu} - P(\delta = 1)^{-1}\right\} \\ & - P(\delta = 1) E\left\{W U_{\theta^*}(X, Y) D_\tau^*(X, Z)^\top\right\} N^{1/2} \hat{\zeta} \\ & + E\left\{\frac{\partial U_{\theta^*}(X, Y)}{\partial \theta}\right\} N^{1/2} \left(\hat{\theta}_{\mathrm{EL3}} - \theta^*\right) + o_p(1). \end{aligned} \tag{62}$$

By using the method of Lagrange multipliers and the Taylor expansion, we obtain

$$N^{1/2} \begin{pmatrix} \hat{\nu} - P(\delta = 1)^{-1} \\ \hat{\zeta} \\ \hat{\tau} - \tau^* \end{pmatrix} = -K^{-1} \begin{pmatrix} N^{-1/2} \sum_{i=1}^{N} (1 - \delta_i W_i) \\ N^{-1/2} \sum_{i=1}^{N} \delta_i W_i D_{\tau^*}^*(X_i, Z_i) \\ \{P(\delta = 1)\}^{-1} \rho \Sigma_1^{-1} N^{1/2} (\tilde{\tau} - \tau^*) \end{pmatrix} + o_p(1),$$

where

$$K = \begin{pmatrix} P(\delta = 1) E(1 - W) & P(\delta = 1) E\{(W-1) D_\tau^*\}^\top & O \\ P(\delta = 1) E\{(W-1) D_\tau^*\} & -P(\delta = 1) E(W D_\tau^{*\otimes 2}) & E(\partial D_\tau^* / \partial \tau) \\ O & E(\partial D_\tau^* / \partial \tau)^\top & P(\delta = 1)^{-1} \rho \Sigma_1^{-1} \end{pmatrix}. \tag{63}$$

The terms in (62) can be simplified to

$$\begin{aligned} & -\begin{pmatrix} P(\delta = 1) E\{(W-1) U_{\theta^*}\} & -P(\delta = 1) E\left\{W U_{\theta^*}(X, Y) D_\tau^{*\top}\right\} & O \end{pmatrix} K^{-1} \\ & = (G_1 \quad G_2 \quad G_3), \end{aligned}$$

where

$$\begin{aligned} G_1 &= \frac{E\{(W-1) U_{\theta^*}\}}{E(W-1)} - E\left(\frac{\partial U_{\theta^*}}{\partial \theta^\top}\right) M E\left(\frac{\partial D_\tau^*}{\partial \tau^\top}\right)^{-1} \frac{E\{(W-1) D_\tau^*\}}{E(W-1)}, \\ G_2 &= -E\left(\frac{\partial U_{\theta^*}}{\partial \theta^\top}\right) M E\left(\frac{\partial D_\tau^*}{\partial \tau^\top}\right)^{-1}, \\ G_3 &= P(\delta = 1) E\left(\frac{\partial U_{\theta^*}}{\partial \theta^\top}\right) M \frac{\Sigma_1}{\rho}, \\ M &= E\left(\phi_{\mathrm{eff}} \eta_{\mathrm{eff}}^\top\right) \left\{\frac{\Sigma_1}{\rho} + E\left(\eta_{\mathrm{eff}}^{\otimes 2}\right)\right\}^{-1}, \end{aligned}$$

and ϕ_{eff} and η_{eff} are the efficient influence functions for θ and τ based on the internal individual data. By the Taylor expansion (62) and (63), we can show that

$$
\begin{aligned}
& -N^{1/2}\left(\hat{\theta}_{\mathrm{EL3}}-\theta^*\right) \\
= & E\left(\frac{\partial U_{\theta^*}}{\partial \theta^{\top}}\right)^{-1} N^{-1/2} \sum_{i=1}^N \delta_i W_i\left(\frac{R_i}{\pi_i^{[1]}} U_{\theta^*}+\frac{\pi_i^{[1]}-R_i}{\pi_i^{[1]}} g_{\theta^*}^{[1]}\right) \\
& +E\left(\frac{\partial U_{\theta^*}}{\partial \theta^{\top}}\right)^{-1} G_1 N^{-1/2} \sum_{i=1}^N(1-\delta_i W_i)+E\left(\frac{\partial U_{\theta^*}}{\partial \theta^{\top}}\right)^{-1} G_2 N^{-1/2} \sum_{i=1}^N \delta_i W_i D_\tau^*(X_i, Z_i) \\
& +E\left(\frac{\partial U_{\theta^*}}{\partial \theta^{\top}}\right)^{-1} G_3 \frac{1}{P(\delta=1)} \rho \Sigma_1^{-1} N^{1/2}\left(\tilde{\tau}-\tau^*\right)+o_p(1) \\
= & E\left(\frac{\partial U_{\theta^*}}{\partial \theta^{\top}}\right)^{-1} N^{-1/2} \sum_{i=1}^N\left\{\delta_i W_i\left(\frac{R_i}{\pi_i^{[1]}} U_{\theta^*}+\frac{\pi_i^{[1]}-R_i}{\pi_i^{[1]}} g_{\theta^*}^{[1]}\right)+(1-\delta_i W_i) \frac{E\left\{(W-1) U_{\theta^*}\right\}}{E(W-1)}\right\} \\
& -M E\left(\frac{\partial D_\tau^*}{\partial \tau^{\top}}\right)^{-1} N^{-1/2} \sum_{i=1}^N(1-\delta_i W_i) \frac{E\left\{(W-1) D_\tau^*\right\}}{E(W-1)} \\
& -M E\left(\frac{\partial D_\tau^*}{\partial \tau^{\top}}\right)^{-1} N^{-1/2} \sum_{i=1}^N \delta_i W_i D_\tau^*(X_i, Z_i)+M N^{1/2}\left(\tilde{\tau}-\tau^*\right)+o_p(1) \\
= & N^{-1/2} \sum_{i=1}^N\left(\phi_{\mathrm{eff}, i}-M \eta_{\mathrm{eff}, i}\right)+M N^{1/2}\left(\tilde{\tau}-\tau^*\right)+o_p(1).
\end{aligned}
$$

Thus, the influence function of the two-step empirical-likelihood estimator in Setting 3 is asymptotically the same. □

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Dr. Kosuke Morikawa, for their invaluable guidance, encouragement, and support throughout my doctoral studies. His insightful comments during seminars are always incredibly educational. They played a significant role in sparking my interest in statistics and guiding my transition from mathematics to this field. Under his guidance, I have developed a stronger interest in the United States, which will certainly play a major role in shaping my future career. He has imparted to me invaluable ways of thinking that are fundamental for a researcher in statistics. I am deeply and profoundly grateful for the extraordinary privilege and good fortune of having been able to study under his guidance.

My gratitude goes to Professor Masayuki Uchida for their unwavering support throughout my doctoral course. Their excellent management of the laboratory fostered an environment where I could concentrate fully on my research. I am particularly thankful for their detailed review of my thesis as the principal reviewer and for providing constructive feedback that significantly enhanced the quality of my work.

I wish to thank Professor Tomoyuki Sugimoto and Professor Joe Suzuki for their valuable comments and suggestions. I am also truly grateful for the many opportunities I received through the dedicated management of the statistics research group.

I am deeply indebted to Professor Yutaka Kano for their invaluable support and comments on the research. I was really pleased with the various discussions on my first paper revision.

Part of this work would not have been possible without Dr. Jongho Im and Mr. Jinung Choi who are my collaborators and stood by me through the long journey of peer review.

I wish to acknowledge Dr. Yoshikazu Terada for his strong passion for mathematics, as our numerous discussions truly deepened my own interest and reaffirmed my interest in mathematics.

I am deeply grateful to Professor Changbao Wu for supervising me during my research stay in Canada. His support not only helped broaden my perspective on pursuing an international research career but also provided me with kind and generous guidance throughout my stay.

Many supports from Professor Satoshi Hattori and Dr. Sho Komukai played a significant role in inspiring me to pursue a path in biostatistics. Their permission to attend the weekly seminars allowed me to learn an incredible amount and I am deeply grateful.

I am truly thankful to the Mathematical Statistics Group at Hiroshima University, where I spent my undergraduate years, for instilling in me the fundamentals of mathematical statistics, which have been invaluable throughout my academic journey.

Finally, I would like to extend my heartfelt thanks to my colleagues, especially, Mr. Kotaro Mizuma, Mr. Hiroki Chiba, and Mr. Wataru Aida, and family, whose encouragement and understanding have been a source of strength and motivation throughout this course.

List of publications

1. Beppu, K., and Choi, J., Morikawa, K. and Im, J. (2024). Identification enhanced generalised linear model estimation with nonignorable missing outcomes. *arXiv* : 2204.10508.

2. Beppu, K. and Morikawa, K. (2024). Verifiable identification condition for nonignorable nonresponse data with categorical instrumental variables. *Statistical Theory and Related Fields*, **8**(1), 40-50.

3. Morikawa, K., Beppu, K. and Aida, W. (2024). Efficient Multiple-Robust Estimation for Nonresponse Data Under Informative Sampling. *arXiv* : 2311.06719.