IDENTIFICATION AND DOUBLY ROBUST ESTIMATION OF DATA MISSING NOT AT RANDOM WITH A SHADOW VARIABLE

BY WANG MIAO∗, ERIC TCHETGEN TCHETGEN† AND ZHI GENG‡

∗, ‡Beijing University, †Harvard University

We consider identification and estimation with an outcome missing not at random (MNAR). We study an identification strategy based on a so-called shadow variable. A shadow variable is assumed to be correlated with the outcome, but independent of the missingness mechanism conditional on the outcome. We give a necessary and sufficient condition for identification of the full data law given a valid shadow variable under MNAR, and also sufficient conditions which are convenient to verify in practice. The conditions are satisfied by many commonly-used models, and thus essentially state that lack of identification is not an issue in many situations. Focusing on estimation of an outcome mean, we describe three semiparametric estimation methods: inverse probability weighting, outcome regression and doubly robust estimation. We evaluate the finite sample performance of these estimators via simulations, and apply them to a China Home Pricing dataset extracted from the China Family Panel Survey (CFPS).

1. Introduction. Methods for missing data have received much attention in statistics and the social sciences. Data are said to be missing at random (MAR) if the missingness mechanism only depends on the observed data, otherwise, data are said to be missing not at random [MNAR, Rubin (1976)]. Considering inference with an outcome subject to missingness, it is well known that the underlying full data law and its functionals are identified under MAR, and corresponding methods to make inference abound, to name a few, likelihood based methods (Dempster, Laird and Rubin, 1977), imputation (Schenker and Welsh, 1988; Rubin, 2004), inverse probability weighting (Horvitz and Thompson, 1952; Robins, Rotnitzky and Zhao, 1994), and doubly robust methods (Van der Laan and Robins, 2003; Bang and Robins, 2005; Tsiatis, 2007). Among them, the doubly robust

∗Supported by the China Scholarship Council.
†Supported by NIH grants AI113251, ES020337 and AI104459.
‡Supported by NSFC 11171365, 11021463 and 10931002.
MSC 2010 subject classifications: Primary 62A01; secondary 62D05, 62G35.
Keywords and phrases: Missing not at random, identification, doubly robust estimation, shadow variable, China home price.
approach is in principle most robust, since it requires correct specification of either a model for the full data law, or of the missing data mechanism, but not necessarily both; while direct likelihood or imputation methods rely on correct specification of the full data law, and likewise inverse probability weighting relies on correct specification of the missing data mechanism. Since doubly robust methods effectively double one’s chances to reduce bias due to model misspecification, such methods have grown in popularity in recent years for estimation with missing data and other forms of coarsening problems, such as encountered in causal inference (Van der Laan and Robins, 2003; Bang and Robins, 2005; Tsiatis, 2007).

Compared to MAR, MNAR is much more challenging. Under MNAR, even parametric models are often non-identifiable (Miao, Ding and Geng, 2014; Wang, Shao and Kim, 2014). However, MNAR is of common occurrence in practice when missingness depends on the missing values even conditional on the observed data. Several authors have studied the problem of identification under MNAR. Among them, Heckman (1979) proposed the so-called Heckman Selection Model, which includes a pair of models for the outcome and the selection process conditional on correlated latent variables to induce an association between the outcome and selection process. Little (1993, 1994) introduced methods based on a pattern-mixture parametrization for incomplete data, which specifies the distribution of the outcome for each missing data pattern separately. Little studied identification of pattern-mixture models by imposing restrictions on unidentifiable parameters across different missing data patterns, for example, setting the parameters of the joint distribution of the data under a given pattern of missingness to be equal to a known function of parameters under a different pattern. Fay (1986) and Ma, Geng and Hu (2003) employed graphical models for the missing data mechanism and studied identification for categorical variables. Rotnitzky, Robins and Scharfstein (1998) and Robins, Rotnitzky and Scharfstein (2000) developed sensitivity analysis methods by assuming that the association between the outcome and the missingness process is fully specified. Das, Newey and Vella (2003) and Tchetgen Tchetgen and Wirth (2013) gave sufficient identification conditions under MNAR in nonparametric and semiparametric regression models with a valid instrumental variable, which is correlated to the missingness process but independent of the outcome in the underlying population.

Identification is sometimes also possible, if a fully observed correlate of the outcome, is known to be independent of the missingness conditional on observed covariates and the outcome of interest. Such a correlate, which we refer to as an shadow variable, is available in many empirical studies. For
example, in a study of mental health of children in Connecticut (Zahner et al., 1992; Ibrahim, Lipsitz and Horton, 2001), researchers were interested in evaluating the prevalence of students sampled from a metropolitan center with abnormal psychopathological status based on their teacher’s assessment, which was subject to missingness. As indicated by Ibrahim, Lipsitz and Horton (2001), a missing teacher report may be related to the teacher’s assessment of the student even upon adjusting for fully observed covariates which included parental status of the household and physical health of the child. Interestingly, a separate parent report on the psychopathology of the child was available for all children in the study. Such a report is likely highly correlated with that of the teacher, however, it is unlikely to be related to the teacher’s response rate conditional on her assessment of the student, in which case, the parental assessment constitutes a valid shadow variable.

Another example we will consider in the application section of the paper concerns the China Home Pricing study. In the survey, residents were asked the current price of their home, which was unknown to some participants. Those who failed to respond, did so out of a lack of awareness about the current market value of their home because of the level of specialized knowledge about the real estate market required in some residential areas. Nonetheless, homeowners typically remembered the original market value of their home when it was either acquired or built. The original price is expected to be highly correlated with the current price, however, it is unlikely to affect a homeowner’s knowledge of the current price conditional on the latter and other covariates, such as geographic location of the home and the travel time to the closest business center. Thus the original price of a home may be selected as a shadow variable for homeowner nonresponse about the current value of his or her home.

Even with a shadow variable, identification is not always guaranteed without an additional assumption. DHaultfoeuille (2010) studied identification with a shadow variable of a regression model with separable error and a nonparametric missing data mechanism, and presented nonparametric estimation methods. Wang, Shao and Kim (2014) studied identification allowing the outcome regression to be less restricted, however with a parametric missing data mechanism, and proposed inverse probability weighted estimation for the mean of the outcome. Zhao and Shao (2014) studied identification of a generalized linear model with unrestricted missing data mechanism, and developed estimation methods based on pseudo-likelihood.

Several methods of estimation initially developed under MAR have recently been extended to handle MNAR problems under suitable conditions. For instance, maximum likelihood estimation (Greenlees, Reece and Zi-
eschang, 1982), inverse probability weighted estimation (Scharfstein, Rotnitzky and Robins, 1999), outcome regression based estimation (Scharfstein and Irizarry, 2003; Vansteelandt, Rotnitzky and Robins, 2007). In contrast, doubly robust estimation methods for data missing not at random are not as well developed. Although Scharfstein and Irizarry (2003) and Vansteelandt, Rotnitzky and Robins (2007) proposed doubly robust estimators for the outcome mean, their approach entailed a sensitivity analysis in which the unidentifiable association between the outcome of interest and the missingness mechanism is assumed to be known, and the outcome mean is estimated over a range of values for the assumed association.

In this paper, we consider identification and estimation with an outcome missing not at random, given a valid shadow variable. The shadow variable approach may be viewed as counterpart to an instrumental variable, and enjoys different identification properties also leading to different methods for estimation. We present a very general identification framework with a shadow variable, considering in turn identification under a selection model parametrization in Section 2, followed by identification under a pattern-mixture parametrization in Section 3. Our results provide for each setting necessary and sufficient conditions for identification of the full data law. In addition, we give general sufficient conditions for identification that are in principle more convenient to verify in nonparametric and semiparametric models. While we establish identification conditions of the joint distribution for the underlying full data, in order to simplify the exposition, we study inference in the context of estimation of the outcome mean, although the methods we describe can be adapted for other functionals without much difficulty.

In Section 4, we model the joint distribution of the outcome, the shadow variable and the missingness process via a pattern-mixture parametrization. The joint distribution is factored into three parts: a baseline outcome model, which models the joint distribution of the outcome and the shadow variable in complete-cases; a baseline missing data mechanism, which models the missingness mechanism at a baseline value of the outcome; and a log odds ratio model, which encodes the association of the outcome and the missingness mechanism. Based on such models, we propose a suite of estimators, that rely on correct specification of some but not necessarily all models. Specifically, we propose an inverse probability weighted estimator, a regression based estimator, and we also propose a doubly robust estimator for the outcome mean. Provided correct specification of the log odds ratio model, the doubly robust estimator is consistent if either the baseline outcome model or the baseline missing data mechanism is correctly specified. In contrast, the inverse probability weighted estimator requires a correct
model for the baseline missingness mechanism, and the outcome regression based estimator requires a correct model for the baseline outcome model. We study the finite sample performance of the estimators in Section 5 via a series of simulations. Simulation results confirm the double robustness property of the proposed doubly robust estimator, and illustrate the bias of the other estimators when a required model is incorrect. In Section 6, we apply the methods to the China Home Pricing example. We conclude in Section 7, and relegate proofs to the Appendix and supplemental discussions to the Supplementary Material.

2. Identification under a selection model parametrization. Throughout the paper, we let $Y$ denote the outcome, $R$ is the missingness indicator with $R = 1$ if $Y$ is observed, otherwise $R = 0$, and $X$ denotes fully observed covariates. Suppose that one has also fully observed a variable $Z$ that satisfies the following conditions of a shadow variable.

**Assumption 1.** (a) $Z \perp R \mid (Y, X)$; (b) $Z \perp Y \mid X$.

The assumption formalizes the idea that $Z$ only affects missingness through its association with $Y$. We use lower-case letters for realized values of corresponding variables, for example, $y$ for a value of the outcome variable $Y$. For simplicity, we omit $X$ in the following, in which case (a) and (b) become $Z \perp R \mid Y$ and $Z \perp Y$. However, all the identification results will hold with covariates by replacing the required conditions with their versions conditional on covariates. The observed data are $n$ identically and independently distributed samples, with some values of $Y$ missing. Thus the observed data are $(Z, RY, R)$. We use the symbol $P$ to denote a joint density function, for instance, $P(z, y, r)$ for the joint density of $(Z, Y, R)$. A person’s contribution to the observed data likelihood is

$$P(z, y, r = 1)^r P(z, r = 0)^{1-r},$$

which represents all of the information contained in the observed data. The observed data likelihood is a functional of the joint distribution of $(Z, Y, R)$, however, given the observed data likelihood, the joint distribution may not be uniquely determined. Let $P_{Z,Y,R}$ denote a given model for the joint distribution of $(Z, Y, R)$, which is a collection of all candidates of $P(z, y, r)$. The joint distribution is said to be identifiable if and only if given the observed data likelihood, the element of $P_{Z,Y,R}$ is uniquely determined.

In this section, we study identification of selection models with a shadow variable. The selection model specifies the missing data mechanism and the
distribution of the outcome separately. Under Assumption 1, we factorize
the joint density as
\[ P(z, y, r) = P(r|y)P(z, y). \]
We first consider a binary data example, and demonstrate identification of
the joint distribution. Then we extend to general cases and give a necessary
and sufficient condition for identification. We also present a sufficient con-
dition which is more straightforward to check in practice. We illustrate the
conditions with a number of prominent examples.

2.1. The binary case. Suppose \( Z \) and \( Y \) are binary. One is only able to
identify the quantities \( P(z, y|r = 1) \), \( P(z|r = 0) \) and \( P(r = 1) \) from the
observed data. These quantities are functions of the unknown parameters
\( P(z, y) \) and \( P(r = 1|y) \). Without imposing any further model assumptions,
we have five unknown parameters, and five independent estimating equa-
tions. Intuitively, we can solve for the parameters from the available equa-
tions to identify them, provided the solution is unique. The approach is
formalized in the following theorem for binary data.

**Theorem 1.** Suppose Assumption 1 holds, then the joint distribution of
\((Z, Y, R)\) is uniquely identified from the observed data \((Z, RY, R)\).

Theorem 1 was first established by Ma, Geng and Hu (2003), and the proof
can be found there. Assumption 1 is indispensable for identification. The
first part of the assumption says that \( Z \) is shadow in the sense that \( Z \) is not
included in the missing data mechanism. It renders the number of unknown
parameters equal to the number of available estimating equations, which is
necessary for identification. The second part, in fact, imposes a full rank
condition on the system of equations, and further guarantees uniqueness of
its solution.

2.2. General cases. For general outcomes and shadow variables, an ex-
plicit rank condition is no longer available. However, as before, identification
remains possible only for a subset of all data generating mechanisms which
we characterize next. In this vein, we let \( \mathcal{P}_{R|Y} \), \( \mathcal{P}_{Y|Z} \) and \( \mathcal{P}_Z \) denote the sets
of laws that satisfy Assumption 1 for \( P(r = 1|y) \), \( P(y|z) \) and \( P(z) \), respec-
tively. Because \( Z \) is fully observed, \( P(z) \) is identified even when \( \mathcal{P}_Z \) includes
all distributions for \( Z \). However, we must further reduce membership for
\( \mathcal{P}_{R|Y} \) or/and \( \mathcal{P}_{Y|Z} \) to guarantee identification. Identification of the latter
models is determined by the relationship between their respective members.
Theorem 2. Under Assumption 1, the joint density \( P(z, y, r) \) is identified from the observed data, if and only if \( \mathcal{P}_{R|Y} \) and \( \mathcal{P}_{Y|Z} \) satisfy the condition: for any two pairs of elements \( P_1(r = 1|y), P_1(y|z) \) and \( P_2(r = 1|y), P_2(y|z) \), their ratios are not equal, that is, with a positive probability

\[
\frac{P_1(y|z)}{P_2(y|z)} \neq \frac{P_2(r = 1|y)}{P_1(r = 1|y)}.
\]

Theorem 2 presents a necessary and sufficient condition for identification of the joint distribution, and thus a sufficient condition for identification of its functionals, such as the outcome mean. It is quite convenient to check the conditions in Theorem 2 for parametric models. We further illustrate it with a binary example.

Example 1. The binary case. We consider a saturated selection model with \( \mathcal{P}_{Y|Z} = \{P(y = 1|z) = \eta_{1z}, z = 0, 1\} \). Then for any two candidates \( P_1(y = 1|z) = \eta_{1z} \) and \( P_2(y = 1|z) = \eta_{2z} \), we have

\[
\frac{P_1(y|z)}{P_2(y|z)} = \frac{\eta_{1z}^y(1 - \eta_{1z})^{1-y}}{\eta_{2z}^y(1 - \eta_{2z})^{1-y}}.
\]

Note the assumption \( Z \perp Y \), implying \( \eta_{1i} \neq \eta_{0i} \) for \( i = 1, 2 \). We establish in the Supplementary Material that the above ratio varies with \( z \) for different candidates of \( P(y|z) \). However, the ratio of any two missing data mechanisms is a function of \( y \). Thus, the condition of Theorem 2 is satisfied, and therefore the joint distribution of \((Z, Y, R)\) is identified.

The conclusion reached in Example 1 is also consistent with Theorem 1. The binary case has a very clear candidate set of models, however, for semiparametric and nonparametric models, explicit expressions for candidate sets are generally not available, so that, the condition of Theorem 2 is not convenient to check. The following corollary provides a more practical approach to check for identification in such models. It follows from the fact that \( P_1(r = 1|y)/P_2(r = 1|y) \) is a function of \( y \), which is usually not equal to \( P_1(y|z)/P_2(y|z) \), a function that varies with \( z \).

Corollary 1. For any two candidates \( P_1(y|z) \) and \( P_2(y|z) \) of \( \mathcal{P}_{Y|Z} \), if the ratio \( P_1(y|z)/P_2(y|z) \) is a constant or varies with \( z \), then the joint distribution of \((Z, Y, R)\) is identifiable under Assumption 1.

Note that we do not specify the missing data mechanism in Corollary 1. Theorem 2 and Corollary 1 are not trivial, since we do have counterexamples
that fail to satisfy the conditions, such as the saturated model of trinary $Y$ with binary $Z$. However, the corollary still allows for many commonly-used models. We consider models for the case of a continuous shadow variable, for instance, the linear regression model:

$$Y = \beta_0 + \beta_1 Z + \varepsilon, \quad \varepsilon \perp \perp Z.$$ 

The model does not specify the distribution of the shadow variable as well of the error term, so the joint distribution is semiparametric with two parameters $(\beta_0, \beta_1)$ and infinite dimensional parameters $P_\varepsilon, P_Z$ and $P(r = 1|y)$. Besides Assumption 1, the missing data mechanism is otherwise nonparametric. Nonetheless, the joint distribution of $(Z, Y, R)$ is identifiable. One can likewise show that identification remains possible in the even larger nonparametric regression model with independent additive error.

(1) $$Y = \mu(Z) + \varepsilon, \quad \varepsilon \perp \perp Z,$$

with $\mu$ unrestricted. Under regularity conditions, we can prove that $P_{Y|Z}$ satisfies the condition of Corollary 1, and thus the joint distribution of $(Z, Y, R)$ is identifiable. The identification of such models can be further strengthened by considering the heteroscedastic model:

(2) $$Y = \mu(Z) + \sigma(Z)\varepsilon, \quad \varepsilon \perp \perp Z,$$

or equivalently

$$P(y|z) = \frac{1}{\sigma(z)}P_\varepsilon \left\{ \frac{y - \mu(z)}{\sigma(z)} \right\},$$

with $\mu, \sigma$ unrestricted, and unknown probability density function $P_\varepsilon$. This class of models is quite common in statistical analysis, such as linear and nonlinear regression. The model is identifiable under MNAR with the help of a shadow variable.

**Theorem 3.** Suppose the heteroscedastic model (2) with a continuous shadow variable, and the density function of the error term $P_\varepsilon$ satisfies the regularity conditions in the Appendix and condition

(a) for some linear and one-to-one mapping $M : P\{(\varepsilon - a)/b\} \mapsto G(t, a, b)$, and any $b, b' > 0$, $(a, b) \neq (a', b')$,

$$\lim_{t \to t_0} G(t, a, b)/G(t, a', b') = 0 \text{ or } \infty \text{ for some } t_0.$$
Then $\mathcal{P}_{Y|Z}$ satisfies the condition of Corollary 1, and thus the joint distribution of $(Z, Y, R)$ is identified from the observed data under Assumption 1 for an otherwise unrestricted model for the missingness process.

Theorem 3 shows identification of a very large class of models. Identification results of model (1) were previously established by DHaultfoeuille (2010), and model (2) allowing heteroscedastic error term can be viewed as a generalization of their results. It is particularly interesting because it essentially states that lack of identification is not an issue in many situations. Condition (a) of Theorem 3 imposed on the distribution of the error term states that the impact of the error term and that of the shadow variable on the outcome is distinguishable to some extent. It is satisfied by many commonly-used models, for instance, the normal error density with $M$ being the identity mapping and $G$ being the density function itself, and the Student-t error density with $M$ being the inverse Fourier transform and $G$ being its characteristic function. For illustration, consider the following simple normal outcome model: let $\mathcal{P}_{Y|Z} = \{N(\gamma_0 + \gamma_1 Z, \sigma_1^2) : \gamma_0, \gamma_1, \sigma_1^2\}$, $\mathcal{P}_Z = \{N(\beta, \sigma_2^2) : \beta, \sigma_2^2\}$. According to Theorem 3, the joint distribution of $(Z, Y, R)$ is identifiable under such outcome model with an unrestricted missing data mechanism. Note however, that one cannot underestimate the central role of the shadow variable in all examples given above. Without such a variable, identification is no longer guaranteed in these examples without imposing an alternative assumption. For instance, let us revisit the normal outcome model in which the shadow variable is no longer available, i.e. Assumption 1 does not hold. Let $\mathcal{P}_Y = \{N(\gamma, \sigma^2) : \gamma, \sigma^2\}$. Without further restrictions on the missing data mechanism, the joint distribution of $(Y, R)$ is not identifiable. Moreover, even if one were to assume a parametric model for the missing data mechanism, identification is not guaranteed. For example, as one assumes the Logistic missing data mechanism, $\mathcal{P}_{R|Y} = \{\logit P(r = 1|y) = \alpha_0 + \alpha_1 y : \alpha_0, \alpha_1\}$, counterexamples to identification are easy to find, see Miao, Ding and Geng (2014) and Wang, Shao and Kim (2014).

3. Identification under a pattern-mixture parametrization. As an alternative to the selection model parametrization, the pattern-mixture parametrization factorizes the joint density of $(Z, Y, R)$ as

$$P(z, y, r) = P(z, y|r)P(r).$$

As pointed by Little (1993, 1994), the parameters of $P(z, y|r = 0)$ are not identifiable in general. However, he showed that identification can be recovered by imposing sufficient restrictions on the parameters of the density
of $Y$ across different missing data patterns. Instead of imposing functional restrictions, here we introduce the log odds ratio function relating $(Z,Y)$ and $R$. In a slight abuse of notation, we denote the log odds ratio function by $OR(y,z)$, which is defined as a measure of the deviation between $P(z,y|r = 1)$ and $P(z,y|r = 0)$:

\begin{equation}
OR(z, y) = \log \frac{P(z, y|r = 0)P(r = 1|y = 0)}{P(z, y|r = 1)P(r = 0|y = 0)}.
\end{equation}

Here $(z = 0, y = 0)$ is a user-specified baseline value of $(Z,Y)$, which may be set to any value in the support of $(Z,Y)$. We note the following equivalent representation under Assumption 1

\begin{equation}
OR(z, y) = \log \frac{P(r = 0|y)P(r = 1|y = 0)}{P(r = 0|y = 0)P(r = 1|y)}.
\end{equation}

Under Assumption 1, the log odds ratio is only a function of $y$, which we therefore denoted $OR(y)$. As shown above, the log odds ratio function also encodes the degree to which the missing data process departs from MAR, which corresponds to $OR(y) = 0$. The joint distribution of $(Z,Y,R)$ is uniquely determined by $OR(y)$, $P(z,y|r)$, and $P(r = 1|y = 0)$:

\begin{equation}
P(z, y, r) = C_1 \exp\{(1 - r)OR(y)\}P(r = 1|y = 0)P(z, y|r = 1) = C_2 \exp\{-rOR(y)\}P(r = 0)P(z, y|r = 0),
\end{equation}

where $C_1, C_2$ are normalizing constants, and $\mathbb{E}[\exp\{OR(y)\}|r = 1] < \infty$. The symbol $\mathbb{E}$ stands for expectation. One can verify that $C_1 = P(r = 1)/P(r = 1|y = 0)$, $C_2 = P(r = 0)/P(r = 0|y = 0)$.

Identification under a pattern-mixture parametrization depends on models for the conditional density of $(Z,Y)$ given $R = 0$ and the log odds ratio. We let $P_{Y|Z,0}$ and $OR_Y$ denote models for $P(y|z, r = 0)$ and $OR(y)$ respectively, that satisfy Assumption 1. We reduce the membership of $P_{Y|Z,0}$ or/and $OR_Y$ to guarantee identification.

**Theorem 4.** Under Assumption 1, the joint density $P(z, y, r)$ is identifiable, if and only if $P_{Y|Z,0}$ and $OR_Y$ satisfy the following condition: for any two pairs of elements $P_1(y|z, r = 0), OR_1(y)$ and $P_2(y|z, r = 0), OR_2(y)$, their ratios are not equal up to the constant $C_{1,2}$, that is, with a positive probability

\[
\frac{P_1(y|z, r = 0)}{P_2(y|z, r = 0)} \neq C_{1,2} \frac{\exp\{OR_1(y)\}}{\exp\{OR_2(y)\}},
\]

where $C_{1,2} = \mathbb{E}[\exp\{-OR_1(y)\}|r = 0]/\mathbb{E}[\exp\{-OR_2(y)\}|r = 0]$. 

The constant $C_{1,2}$ is related to the log odds ratio functions $OR_1(y)$ and $OR_2(y)$. An immediate sufficient condition is that the ratios presented in the theorem never equal up to any constant, and a corollary similar to Corollary 1 follows.

**Corollary 2.** If the ratio $P_1(y|z, r = 0)/P_2(y|z, r = 0)$ is a constant or varies with $z$ for any two elements of $P_{Y|Z,0}$, then the joint distribution of $(Z,Y,R)$ is identified under Assumption 1.

Consider again the binary data case. Then we have the following result.

**Example 2.** We consider a saturated pattern-mixture model $P_{Y|Z,0} = \{P(y = 1|z, r = 0) = \theta_z, z = 0, 1\}$. Then for any two candidates $P_1(y = 1|z, r = 0) = \theta_{1z}$ and $P_2(y = 1|z, r = 0) = \theta_{2z}$, we have

$$\frac{P_1(y|z, r = 0)}{P_2(y|z, r = 0)} = \frac{\theta_{1z}^y(1 - \theta_{1z})^{1-y}}{\theta_{2z}^y(1 - \theta_{2z})^{1-y}}.$$ 

Note the assumption $Z \perp Y$, we prove in the Supplementary Material that $\theta_{1i} \neq \theta_{0i}$ for $i = 1, 2$, and that the above ratio must vary with $z$ for different candidates of $P(y|z, r = 0)$. Thus, the condition of Theorem 4 is satisfied, and therefore the joint distribution of $(Z,Y,R)$ is identifiable.

We can also consider the general heteroscedastic regression model with a continuous shadow variable:

$$P(y|z, r) = \frac{1}{\sigma_r(z)} P_{z,r} \left\{ \frac{y - \mu_r(z)}{\sigma_r(z)} \right\}, \quad r = 0, 1,$$

with $\mu_r$ and $\sigma_r$ unrestricted, and $P_{z,r}, r = 0, 1$ are unrestricted density functions. Such models provide a sufficient condition for identification of the joint distribution of $(Z,Y,R)$.

**Theorem 5.** Suppose either $P(y|z, r = 1)$ or $P(y|z, r = 0)$ follows model (5) with the corresponding density function $P_{z,r}$ satisfying condition (a) of Theorem 3 as well as certain regularity conditions given in the Appendix. Then $P_{Y|Z,0}$ satisfies the condition of Corollary 2, and thus the joint distribution of $(Z,Y,R)$ is identified.

Little (1993, 1994) studied identification of pattern-mixture models under various restrictions on the parameters of different patterns. For instance, the missing completely at random [MCAR, Little (1993); Little and Rubin (2002)] mechanism requires identical patterns: $P(z, y|r = 1) = P(z, y|r = 0)$. 


The parameter restriction approach may be particularly useful for parametric models. But it requires precise knowledge of the data generating process, and may be more challenging to impose such restrictions in semiparametric and nonparametric models with infinitely dimensional parameters. However, the results we give above provide an alternative strategy to deal with such models by characterizing the largest class of models that are identifiable, which subsumes the approach of Little (1993). The proposed approach also enjoys the benefit that it is explicit about the key role played by the shadow outcome for identification, without which Theorem 5 fails to hold in general.

4. Estimation. In this section, we mainly focus on inference about the outcome mean, denoted by \( \mu = \mathbb{E}(Y) \), based on the pattern-mixture parametrization. We describe several strategies for estimation, each depending on different modeling assumptions. We take covariates into consideration, and model the joint distribution of \((Z, Y, R)\) conditional on \(X\) in three parts: the log odds ratio model

\[
OR(y|x) = \log \frac{P(z, y|r = 0, x)P(z = 0, y = 0|r = 1, x)}{P(z, y|r = 1, x)P(z = 0, y = 0|r = 0, x)},
\]

the baseline outcome model \(P(z, y|r = 1, x)\), and the baseline missing data mechanism \(P(r = 1|y = 0, x)\). The log odds ratio function does not depend on \(z\) by a similar argument to the case without covariates. We have the following parametrization of the joint density of \((Z, Y, R)\) conditional on \(X\),

\[
P(z, y, r|x) = c(x) \exp\{(1 - r)OR(y|x)\}P(r|y = 0, x)P(z, y|r = 1, x),
\]

with \(\mathbb{E}[\exp\{OR(y|x)\}|r = 1, x] < \infty\), and \(c(x)\) being the normalizing constant given \(x\) making the right hand side a well defined density function. We can express the missing data mechanism and the outcome model for the two patterns as functionals of \(P(z, y|r = 1, x)\), \(P(r = 1|y = 0, x)\) and \(OR(y|x)\).

\[
P(r = 1|y, x) = \frac{P(r = 1|y = 0, x)}{P(r = 1|y = 0, x) + \exp\{OR(y|x)\}P(r = 0|y = 0, x)},
\]

\[
P(z, y|r = 0, x) = \frac{\exp\{OR(y|x)\}P(z, y|r = 1, x)}{\mathbb{E}[\exp\{OR(Y|x)\}|r = 1, x]},
\]

\[
\mathbb{E}(Y|r = 0, x) = \frac{\mathbb{E}[\exp\{OR(Y|x)\}Y|r = 1, x]}{\mathbb{E}[\exp\{OR(Y|x)\}|r = 1, x]}.
\]

These equalities formulate the main parametrization we shall use throughout for estimation. They are straightforward to verify. We specify parametric models for the three parts separately: the baseline missing data mechanism \(P(r = 1|y = 0, x; \alpha)\), the baseline outcome model \(P(z, y|x, r = 1; \beta)\), and the log odds ratio model \(OR(y|x; \gamma)\).
Suppose that the log odds ratio model is correctly specified. First, we describe an inverse probability weighted estimator and an outcome regression based estimator, which require correct specification of the baseline missing data mechanism or the baseline outcome model, respectively. We then develop a doubly robust estimator, which is consistent if either the baseline missing data mechanism or the baseline outcome model is correct, but not necessarily both.

4.1. Inverse probability weighted estimator. The inverse probability weighted estimator relies on the propensity score model \( P(r = 1|y, x; \alpha, \gamma) \), which is determined by \( P(r = 1|y = 0, x; \alpha) \) and \( OR(y|x; \gamma) \) as shown in equation (8). We define the inverse of the propensity score model as \( \hat{w}(y, x; \hat{\alpha}, \hat{\gamma}) = 1/P(r = 1|y, x; \hat{\alpha}, \hat{\gamma}) \), and solve the equation for \( \hat{\alpha} \) and \( \hat{\gamma} \).

\[
P \left[ \{ w(X, Y; \hat{\alpha}, \hat{\gamma}) R - 1\} h(X, Z) \right] = 0,
\]

with a user-specified vector function \( h \), for example, \( h(X, Z) = (X^T, Z^T) \). The symbol \( P \) stands for sample mean. One may decide to allow the dimension of \( h \) to exceed that of \( (\alpha, \gamma) \), in which case one may adopt the Generalized Method of Moments (Hall, 2005). The proposed inverse probability weighted estimator for the mean of the outcome is

\[
\hat{\mu}_{ipw} = P \left\{ w(X, Y; \hat{\alpha}, \hat{\gamma}_{ipw}) RY \right\}.
\]

This estimator was previously described by Wang, Shao and Kim (2014).

4.2. Outcome regression based estimator. As an alternative to inverse probability weighting, we can estimate the mean of the outcome by combining the baseline outcome model and the log odds ratio model. We first obtain a consistent estimate for the baseline outcome model \( P(z, y|r = 1, x; \hat{\beta}) \), such as the maximum likelihood estimator, and then propose to solve the following estimating equation for \( \hat{\gamma}_{reg} \)

\[
P \left[ (1 - R) \left\{ h(Z, X) - \mathbb{E} \left[ h(Z, X) | r = 0, X; \hat{\beta}, \hat{\gamma}_{reg} \right] \right\} \right] = 0,
\]

with previously obtained \( \hat{\beta} \) and a user-specified vector function \( h \). The conditional expectation is evaluated under the conditional density \( P(z|r = 0, x, \hat{\beta}, \hat{\gamma}_{reg}) \), which is determined by \( P(z, y|r = 1, x; \hat{\beta}) \) and \( OR(y|x; \hat{\gamma}_{reg}) \) as in equation (9). The Generalized Method of Moments may also be used here whenever the dimension of \( h \) exceeds that of \( \gamma \). We then obtain an estimator \( \mathbb{E}(Y|r = 0, x; \hat{\beta}, \hat{\gamma}_{reg}) \) from equation (10) for the outcome conditional mean.
among the subset of the population with missing outcome, i.e. \( \mathbb{E}(Y | r = 0, x) \). The regression based estimator is finally given by

\[
\hat{\mu}_{\text{reg}} = \mathbb{P} \left\{ (1 - R) \mathbb{E} (Y | r = 0, x; \hat{\beta}, \hat{\gamma}_{\text{reg}}) + R \mathbb{E} (Y | r = 1, x; \hat{\beta}) \right\}.
\]

Note that an alternative outcome regression-based estimator of \( \mu \) can be obtained by substituting \( Y \) for \( \mathbb{E} (Y | r = 1, x; \hat{\beta}) \).

4.3. **Doubly robust estimator.** The proposed doubly robust estimator requires a doubly robust estimator \( \hat{\gamma}_{\text{dr}} \) for the log odds ratio parameter \( \gamma \), which solves the estimating equation

\[
(13) \quad \mathbb{P} \left\{ w (X, Y; \hat{\alpha}, \hat{\gamma}_{\text{dr}}) R - 1 \right\} \{ h (Z, X) - \mathbb{E} [h (Z, X) | r = 0, X; \hat{\beta}, \hat{\gamma}_{\text{dr}}] \} = 0,
\]

with previously obtained \( \hat{\alpha}, \hat{\beta} \) and a user-specified vector function \( h \). We then obtain an estimator \( \mathbb{E} (Y | r = 0, x; \hat{\beta}, \hat{\gamma}_{\text{dr}}) \) from equation (10) for \( \mathbb{E} (Y | r = 0, x) \). Finally, the proposed doubly robust estimator for the mean of the outcome is

\[
\hat{\mu}_{\text{dr}} = \mathbb{P} \left\{ w (X, Y; \hat{\alpha}, \hat{\gamma}_{\text{dr}}) R \left\{ Y - \mathbb{E} \left[ Y | r = 0, X; \hat{\beta}, \hat{\gamma}_{\text{dr}} \right] \right\} \right. \]
\[
+ \left. \mathbb{E} \left\{ Y | r = 0, X; \hat{\beta}, \hat{\gamma}_{\text{dr}} \right\} \right\}.
\]

The theorem below summarizes the consistency of the estimators proposed above.

**Theorem 6.** Suppose the joint distribution of \((Z, Y, R)\) conditional on \(X\) satisfies the identification condition in Theorem 4; the estimating equations (11), (12), (13) have a unique solution; and the log odds ratio model \( \text{OR}(y | x; \gamma) \) is correctly specified. Then we have

(a) if \( P(r = 1 | y = 0, x; \alpha) \) is correctly specified, then the inverse probability weighted estimator \( \hat{\mu}_{\text{ipw}} \) is consistent;

(b) if \( P(z, y | r = 1, x; \beta) \) is correctly specified, and \( \hat{\beta} \) is consistent, then the outcome regression based estimator \( \hat{\mu}_{\text{reg}} \) is consistent;

(c) if at least one of conditions (a) and (b) holds, then the doubly robust estimators \( \hat{\gamma}_{\text{dr}} \) and \( \hat{\mu}_{\text{dr}} \) are consistent.

The estimators are also asymptotically normal under standard conditions (Hall, 2005). In the Supplemental Material, we derive nonparametric estimators of the asymptotic variance of each of these estimators, which are consistent even under model mis-specification.
The log odds ratio function $OR(y|x)$ plays a central role for estimation under all three proposed strategies, as they all rely on a correct log odds ratio model. This is not entirely surprising, since as previously mentioned, the log odds ratio encodes the degree to which the outcome and the missing data process are correlated. Therefore, in order to estimate a population functional of $Y$, one must first be able to account for its association with the missing data process. Note also that previous doubly robust estimators for missing data have assumed that this log odds ratio is known exactly, either to be equal to the null value of 0 under MAR (Bang and Robins, 2005; Tsiatis, 2007; Van der Laan and Robins, 2003), or to be of a known functional form with no unknown parameter as in Vansteelandt, Rotnitzky and Robins (2007). Without a shadow variable, the log odds ratio function cannot generally be identified in the latter setting if outcome data are missing not at random. Therefore, we have in fact developed a general strategy to relax these previous stringent assumptions. The doubly robust estimator provides us with a second chance to correct the bias due to possible mis-specification of either the baseline outcome model or the baseline missing data mechanism. We have shown that given a correct model for the log odds ratio function, one can be doubly robust both in estimating the association between the outcome and the missingness process, and the outcome mean. However, if both baseline models are incorrect, the doubly robust estimator will generally also be biased.

5. Simulations. In this section, we study the finite sample performance of the proposed methods via simulations. We consider continuous and binary outcomes in turn, however, to save space we relegate simulation results for a binary outcome to the Supplemental Material.

First, we generate a covariate $X \sim N(0, 1)$, and then generate data from the conditional density (7) with a linear log odds ratio model: $OR(y|x) = -\gamma y$, bivariate normal baseline outcome model and Logistic baseline missing data mechanism. We consider two choices for the baseline outcome model: $Y|\tau = 1, x \sim N(\beta_{10} + \beta_{11}(0.5x + 0.2x^2), 1)$ with $Z|y, x \sim N(\beta_{20} + \beta_{21}(2x + x^2) + \beta_{22}y, 1)$; and $Y|\tau = 1, x \sim N(\beta_{10} + \beta_{11}x, 1)$ with $Z|y, x \sim N(\beta_{20} + \beta_{21}x^2 + \beta_{22}y, 1)$; and we consider two choices for the baseline missing data mechanism: logit $P(\tau = 1|y = 0, x) = \alpha_0 + \alpha_1(x - 0.5x^2)$, and logit $P(\tau = 1|y = 0, x) = \alpha_0 + \alpha_1x$. The models are identifiable according to Theorem 5. Parameter values are set equal to $(\alpha_0, \alpha_1) = (0.8, 0.5), (\beta_{10}, \beta_{11}) = (0.5, 0.5), (\beta_{20}, \beta_{21}, \beta_{22}) = (-0.5, 0.5, 1)$, and $\gamma = -0.5$. For these settings, the missing data proportions are between 40% and 45%. We generate data from the four combinations of the baseline outcome model and missing data mechanism,
but always employ a simpler model for estimation: $Y | r = 1, x \sim N(\beta_{10} + \beta_{11} x, 1)$ with $Z | y, x \sim N(\beta_{20} + \beta_{21} x^2 + \beta_{22} y, 1)$, and logit $P(r = 1 | y = 0, x) = \alpha_0 + \alpha_1 x$. We also consider a naive estimator assuming MAR obtained via linear regression based on complete data: $E(Y | x, z) = E(Y | r = 1, x, z) = \beta_{00} + \beta_{01} x + \beta_{02} z$. We simulate 1000 replicates under 500 and 1500 sample sizes for each combination and summarize the results with boxplots.

Figure 1 presents the results for the outcome mean, Figure 2 presents the results for the parameter of the log odds ratio model. Table 1 shows coverage probability of the 0.95 confidence interval estimated with the method in the Supplementary Material. In (i) of Figure 1, the baseline missing data mechanism is incorrect but the baseline outcome model is correct. As a result, the outcome regression based estimator works well and has an appropriate coverage probability, but the inverse probability weighted estimator has very large bias and coverage probability well below the nominal level. It is not consistent as the sample size increases. In (ii), the baseline missing data mechanism is correct but the baseline outcome model is incorrect. The inverse probability weighted estimator has small bias and has an approximate 0.95 coverage probability, but the outcome regression based estimator is biased. However, in (i) and (ii), the doubly robust estimator performs the best with smaller bias and approximate 0.95 coverage probability, and it is consistent as sample size increases. In (iii), both models are correct, and all proposed estimators have similar performance and are all consistent. In (iv), neither of the two models is correct, but the doubly robust estimator has smaller bias than others. We also observe that as expected, the naive estimator assuming MAR is significantly biased in all cases. The performance of the estimators for the log odds ratio model is similar to the estimators for the mean. The results show robustness of the doubly robust estimator. As a conclusion, we recommend the doubly robust approach for inference about the mean parameter as well as to evaluate the magnitude of selection bias.

6. China Home Pricing example. The dataset is extracted from China Family Panel Studies (CFPS). The CFPS project is conducted by Institute of Social Science Survey of Peking University in Beijing since 2008. The survey data contain information on various aspects of the household, such as socioeconomic information and education information of the households. Details of the survey can be found at http://www.isss.edu.cn/cfps/EN/. The dataset we used consists of 3126 households from Liaoning, Shanghai, Henan, Guangdong and Yunnan province of China. The provinces are located to the east of the Heihe-Tengchong line, a line drawn from the town of Heihe in the northeast province of Heilongjiang to Tengchong in the
MNAR WITH A SHADOW VARIABLE

(i) FT

(ii) TF

(iii) TT

(iv) FF

Fig 1: Boxplots of the estimators for the mean of a normal outcome.

(i) FT

(ii) TF

(iii) TT

(iv) FF

Fig 2: Boxplots of the estimators for log odds ratio model of the normal outcome example.

Note for Fig 1 and 2: data are analyzed with four methods: doubly robust estimation (DR), inverse probability weighting (IPW), regression based estimation (REG), and linear regression based on complete data (CMP). In each boxplot, white boxes are for sample size 500, and gray ones for 1500. The horizontal line marks the true value of the parameter.
Table 1

|       | \( \mu \) |   | \( \gamma \) |   |   |   |   |   |
|-------|-----------|---|-------------|---|---|---|---|---|
|       | DR        | IPW | REG        | DR | IPW | REG |
| FT    | 0.953     | 0.748 | 0.942     | 0.960 | 0.771 | 0.956 |
|       | 0.951     | 0.359 | 0.944     | 0.940 | 0.374 | 0.960 |
| TF    | 0.944     | 0.944 | 0.392     | 0.957 | 0.952 | 0.188 |
|       | 0.959     | 0.943 | 0.013     | 0.955 | 0.946 | 0.000 |
| TT    | 0.949     | 0.944 | 0.940     | 0.952 | 0.954 | 0.954 |
|       | 0.947     | 0.943 | 0.934     | 0.948 | 0.952 | 0.955 |
| FF    | 0.762     | 0.760 | 0.235     | 0.837 | 0.721 | 0.287 |
|       | 0.441     | 0.410 | 0.005     | 0.661 | 0.247 | 0.007 |

Note: Coverage probability of the proposed estimators under four simulation situations: FT stands for incorrectly specified baseline missing data mechanism and correctly specified baseline outcome model, and the other three situations are similarly defined. The variance and confidence interval estimation method can be found in the Supplementary Material. The result of each situation includes two rows, of which the first row stands for sample size 500, and the second for 1500.

southwest province of Yunnan. The Heihe-Tengchong line divides the area of China in half. But only 6% of the population lives in the west; 94% of the population lives in the eastern half of the country [page 19, Naughton (2007)]. Shanghai and Guangdong are located in the southeast coast. They are two most developed provinces of China, but are suffering from congestion, pollution and extortionate home prices due to dense population. The other three provinces, Liaoning located in the northeastern, Henan located in the middle, Yunnan located in the southwest of this area, are three less developed provinces, and the Chinese government is committed to accelerating development of this area and improving quality life of its residents. Against the background of booming real estate, home price is a central issue for both the residents and the government. In the survey, the variable of interest is log of current home price (\( \text{lhmpr} \), in \( 10^4 \) RMB yuan). Homeowners were asked the current price of their house, to which 587(18.8%) householders responded “don’t know”, and 9(3%) refused to respond. Therefore they are treated as missing values in our analysis. The dataset contains completely observed covariates: \( \text{prov} \) (1, developed province, 0 less developed province), \( \text{urban} \) (1, urban, 0 rural household), \( \text{ltm} \) (log of travel time to the nearest business center, in minutes), \( \text{lsiz} \) (log of house building area, in square meters), \( \text{lfmsz} \) (log of family size, i.e. number of families living together), \( \text{flr} \) (which floor the family live in), \( \text{lfminc} \) (log of last year family income, in RMB yuan), \( \text{recons} \) (ever experienced demolishment and reconstruction),
and $\text{loripr}$ (log of original home price, in $10^4$ RMB yuan).

In developed areas and big cities, besides building costs, home prices are also affected by many factors: location issues such as bus stops and business center, property management such as maintenance and garden care, the strength of the competition, and renovation potential. As a result, home prices are relatively high and valuation is complicated due to complexities of real estate market. Home price valuation usually relies on dedicated agencies who have specialized knowledge and rich experience with real estate transactions. A householder is seldom well aware of the valuation of their home. However, in less developed areas, real estate market is pristine and slow-growing. There are much fewer factors that affect home prices. As a result, home prices tend to be lower and valuation depends mostly on building costs. Valuation does not need much specialized knowledge, and in fact, transactions and valuation in these areas seldom rely on appraisal companies. Therefore, nonresponse of $\text{lhmpr}$ is likely related to the current value of the home, and the missingness is likely not at random. The original price of the house is related to the current price, however, we expect that the original price is independent of nonresponse conditional on the current value of the house and fully observed covariates. In other words, we assume that two homeowners with houses of equal current value and common covariates do not differ in their probability of nonresponse even if the original purchase price of their respective homes differs. Therefore, we use $\text{loripr}$ as a shadow variable.

Let $X$ denote all other covariates including 1 for the intercept, we assume a linear log odds ratio model: $\text{OR}(\text{lhmpr}|x) = -\gamma \text{lhmpr}$, a Logistic baseline missing data mechanism:

$$\text{logit } P(r = 1|x, \text{lhmpr} = 0) = x\alpha,$$

and a bivariate baseline outcome model:

$$\mathbb{E}(\text{lhmpr}|r = 1, x) = x\beta_1,$$
$$\mathbb{E}(\text{loripr}|r = 1, x, \text{lhmpr}) = x\beta_{21} + \beta_{22}\text{lhmpr}.$$ 

We summarize the results of mean of $\text{lhmpr}$ and the log odds ratio in Table 2, and Table A.2 summarizing model fits for all baseline models is relegated to the Supplementary Material. We first apply standard estimation methods assuming missing at random ($R_{\text{lhmpr}}|X, \text{loripr}$) to the data. From Table 2, the standard regression estimate for mean and 95% confidence interval (in bracket) of $\text{lhmpr}$ is $2.693(2.637, 2.749)$ using a linear regression, and the standard inverse probability weighted estimate is
2.694(2.638, 2.751) using a Logistic propensity score model. Details of standard estimates can be found in the Supplementary Material. However, the estimates using the proposed methods are somewhat different from those using standard methods. The estimate is 2.595(2.530, 2.659) using the doubly robust approach, 2.586(2.518, 2.655) using regression based estimation, and 2.599(2.533, 2.665) using inverse probability weighting; the corresponding estimates for the log odds ratio parameter $\gamma$ are 0.497(0.329, 0.664), 0.745(0.430, 1.060) and 0.413(0.241, 0.586). In the above, the three proposed methods produce similar point and interval estimates for the outcome mean. None of the three confidence intervals includes point estimates from standard methods, and none of the three point estimates falls in the confidence intervals from standard methods. Doubly robust estimation and inverse probability weighted estimation also produce similar point and interval estimates of the log odds ratio parameter $\gamma$. Although regression based estimation produces a slightly higher result, all three confidence intervals of $\gamma$ exclude 0, providing significant empirical evidence of selection bias due to the outcome missing not at random. There is also substantial empirical evidence that homeowners with houses valued at lower prices are less likely to respond to the survey.

| Method  | Mean ($\mu$)     | Log odds ratio ($\gamma$)     |
|---------|------------------|-------------------------------|
| DR      | 2.595(2.530, 2.659) | 0.497(0.329, 0.664)          |
| REG     | 2.586(2.518, 2.655) | 0.745(0.430, 1.060)          |
| IPW     | 2.599(2.533, 2.665) | 0.413(0.241, 0.586)          |
| CMP     | 2.693(2.637, 2.749) |                               |
| marIPW  | 2.694(2.638, 2.751) |                               |

Note: Point estimates and 95% confidence intervals of the outcome mean and log odds ratio parameter. CMP and marIPW stand for standard regression estimation and standard inverse probability weighted estimation.

7. Discussion. The instrumental variable approach is a popular and fairly well developed method for identification of data missing not at random (Das, Newey and Vella, 2003; Tchetgen Tchetgen and Wirth, 2013). In contrast, although recently considered by DHaultfoeuille (2010); Wang, Shao and Kim (2014); Zhao and Shao (2014) in a series of seminal papers, the use of a shadow variable as an alternative identification strategy in the context of data missing not at random is currently not well established in the missing data literature. In this paper, we further explore the use of such a variable, and we establish a general framework for identification, together with a suite
of semiparametric estimators, including a doubly robust approach. The basic idea of a shadow variable approach is to reduce the unidentifiable parameters or candidate models. It is fairly straightforward to illustrate for simple situations but becomes more involved and less convenient for larger, more complicated models such as semiparametric or nonparametric models. We presented necessary and sufficient conditions for identification under a selection model parametrization, and a pattern-mixture parametrization. One may choose to verify the conditions according to which parametrization may be most convenient or appropriate in a given application. As we observed that the sufficient conditions we give are satisfied by many commonly-used models, lack of identification is not an issue in many situations with a shadow variable. We also developed semiparametric estimation methods that extend analogous methods available when data are missing at random.

A few remaining open questions are of significance but not within the scope of the paper. First, the choice of the function \( h \) indexing various estimating equations is directly related to the efficiency of the resulting estimators. Modern semiparametric efficiency theory may be used to identify an optimal choice for such index functions, and therefore to construct a semiparametric locally efficient estimator (Bickel et al., 1993; Van der Vaart, 2000). Second, one may note that in the construction of our doubly robust estimator, the shadow variable was not included as predictor for the outcome model. However, the shadow variable may in fact be treated as any ordinary predictor for the purpose of modeling the outcome, and the efficiency-robustness trade-off of incorporating such a variable in the outcome model remains to be investigated.

The proposed identification and doubly robust estimation methods have potential application in other related topics. The methods can be applied to longitudinal data analysis, which is often subject to dropout or missing data. Their potential use for such more general settings is the topic of future study.

APPENDIX A: PROOFS

Proof of Theorem 2

Proof. The proof is based on contradiction. First, we note that all the quantities we can identify from the observed data are \( P(z, y, r = 1) \) and \( P(z) \). Suppose we have two candidate models satisfying the same observed quantities:

\[
P_1(z, y, r = 1) = P_2(z, y, r = 1), \quad P_1(z) = P_2(z) \text{ almost surely.}
\]
By the shadow variable assumption, we have the factorization for the joint density:

\[ P_i(z, y, r) = P_i(r|y)P_i(y|z)P_i(z), \quad i = 1, 2. \]

So from the above identities we have

\[ P_1(r = 1|y)P_1(y|z) = P_2(r = 1|y)P_2(y|z), \]

and equivalently

\[ \frac{P_1(r = 1|y)}{P_2(r = 1|y)} = \frac{P_2(y|z)}{P_1(y|z)} \text{ almost surely.} \]

The equation contradicts the condition that the candidates of \( P_{R|Y} \) and \( P_{Y|Z} \) make the ratios unequal. So, that unequal ratios is equivalent to the impos-

sibility of two sets of candidates satisfying the same observed quantities, i.e. identifica-
tion of the joint distribution.

The proof of Theorem 3 relies on the following lemma.

**Lemma 1.** If any element of \( P_{Y|Z} \) satisfies:

\[ \int P(y|z)h(y)dy = 0 \text{ for any } z \Rightarrow h = 0, \tag{14} \]

then the ratio of any two elements varies with \( z \), i.e. for any \( P_1(y|z), P_2(y|z) \in P_{Y|Z}, P_1(y|z)/P_2(y|z) \) varies with \( z \).

**Proof.** We prove the lemma by contradiction. Suppose we have two can-
didates \( P_1(y|z) \) and \( P_2(y|z) \) that \( P_1(y|z)/P_2(y|z) = \tilde{h}(y) \) for some function \( \tilde{h} \). Then we have

\[ \int P_2(y|z)dy = \int P_1(y|z)dy = \int P_2(y|z)\tilde{h}(y)dy = 1, \]

and thus \( \int P_2(y|z)(\tilde{h}(y) - 1)dy = 0 \). Since \( \tilde{h} - 1 \neq 0 \), the above equation contradicts condition (14). As a result, the ratio of any two elements of \( P_{Y|Z} \) must vary with \( z \).

Condition (14) is called *completeness* in Hu and Shiu (2011) and DHault-
foeuille (2010). Before we prove Theorem 3, we introduce the following lemma about completeness, which is part of Lemma 4 of Hu and Shiu (2011).
Lemma 2. For every $z$, $P(y|z) \in L^2$, and suppose that there exists a point $z_0$ with its open neighborhood $U(z_0)$ such that

(i) the characteristic function $\phi_{z_0}(t)$ of $P(y|z_0)$ satisfies $0 < |\phi_{z_0}(t)| < C \exp(-\delta |t|)$ for $t \in \mathbb{R}$ and some constants $C, \delta > 0$;

(ii) $\partial P(y|z)/\partial z$ for all $z \in U(z_0)$ and $\partial P(y|z_0)/\partial y$ are in $L^2$;

(iii) there exists a sequence $\{z_k \in U(z_0) : k = 1, 2, \ldots\}$ converging to $z_0$ such that any finite subsequence $P(y|z_{k_i})$ is linearly independent, i.e.

$$\sum_{i=1}^{I} a_i P(y|z_{k_i}) = 0 \text{ implies } a_i = 0, \text{ for all } i = 1, 2, \ldots, I.$$

Then $P(y|z)$ is complete in $L^2$, i.e. condition (14) holds.

We prove Theorem 3 under the following regularity conditions for the density function of the error term:

(1) the characteristic functions $\phi(t)$ of $P(v)$ satisfies $0 < |\phi(t)| < C \exp(-\delta |t|)$ for $t \in \mathbb{R}$ and some constants $C, \delta > 0$;

(2) $\mu(z)$, $\sigma(z)$ and $P(v)$ are continuously differentiable, and $\int_{-\infty}^{+\infty} |v\partial P(v)/\partial v|^2 dv$ is finite.

Proof of Theorem 3

Proof. From Lemma 1, in order to prove Theorem 3, we only need to prove that condition (14) holds under the regularity conditions above and condition (a) of Theorem 3.

Without loss of generality, we choose $z_0$ that $\sigma(z_0) = 1$. Since $\mu$ is continuous with $\partial \mu(z)/\partial z \not= 0$ for $z = z_0$, there exists a distinct sequence $\{z_k \subset U(z_0) : k = 1, 2, 3, \ldots\}$ that $\mu(z_k)$ converges to $\mu(z_0)$. Condition (i) of Lemma 2 is satisfied under regularity condition (1).

We then check that $P(y|z) \in L^2$ for every $z$. We have

$$\int_{y} |P(y|z)|^2 dy = \int_{y} \left| \frac{1}{\sigma(z)} P_{\varepsilon} \left\{ \frac{y - \mu(z)}{\sigma(z)} \right\} \right|^2 dy = \int_{\varepsilon} \frac{1}{\sigma(\varepsilon)} |P_{\varepsilon}(\varepsilon)|^2 d\varepsilon.$$

Since $|\phi(t)| < C \exp(-\delta |t|)$, we have $\int_{t} |\phi(t)|^2 dt < +\infty$, and thus $\int_{\varepsilon} |P_{\varepsilon}(\varepsilon)|^2 d\varepsilon < +\infty$. Note that $\sigma(z) > 0$, we have $P(y|z) \in L^2$ for all $z$. From the regularity condition (2), one can straightforward verify that $\partial P(y|z)/\partial z$ for all $z \in U(z_0)$ and $\partial P(y|z_0)/\partial y$ are in $L^2$.

We verify condition (iii) of Lemma 2 under condition (a) of Theorem 3. Without loss of generality, we assume that for $k = 1, 2, \ldots, I < +\infty$, and
\{a_k \in \mathbb{R} : k = 1, 2, \ldots, I \}
\begin{equation}
  a_1 P(y|z_1) + a_2 P(y|z_2) + \cdots + a_I P(y|z_I) = 0.
\end{equation}

Since the mapping \( M \) is linear and one-to-one, we have
\begin{equation}
  a_1 G(t|z_1) + a_2 G(t|z_2) + \cdots + a_I G(t|z_I) = 0,
\end{equation}
with \( G(t|z_i) = G\{t, \mu(z_i), \sigma(z_i)\} \), and thus for \( t \in \{t : G(t|z_1) \neq 0\} \)
\begin{equation}
  a_1 + a_2 \frac{G(t|z_2)}{G(t|z_1)} + \cdots + a_I \frac{G(t|z_I)}{G(t|z_1)} = 0,
\end{equation}

Note condition (a) of Theorem 3, we assume without loss of generality that
\begin{equation}
  \lim_{t \to t_0} G(t|z_i) = 0, \quad i = 2, \ldots, I,
\end{equation}
then we must have \( a_1 = 0 \). In the same vein, we can prove that \( a_i = 0 \) for \( i = 2, \ldots, I \), which means \( \{P(y|z_i) : i = 1 \ldots, I\} \) is linearly independent. Therefore condition (iii) of Lemma 2 holds under condition (a) Theorem 3.

Finally, conditions (i), (ii) and (iii) hold under the regularity conditions (1), (2) and condition (a) of Theorem 3. Thus, condition (14) holds. \( \square \)

**Proof of Theorem 4**

**Proof.** The proof is by contradiction. We note that the quantities we can identify from the observed data are \( P(z, y, r = 1) \) and \( P(z|r = 0) \). Suppose we have two candidate models satisfying the same observed quantities:
\begin{align*}
P_1(z, y, r = 1) &= P_2(z, y, r = 1), \quad P_1(z|r = 0) = P_2(z|r = 0) \text{ almost surely.}
\end{align*}

By the pattern-mixture parametrization (4), we have
\begin{align*}
P_i(z, y, r) &= C_{2i} \exp\{-OR_i(y)\} P_i(r|y = 0)P_i(z, y|r = 0), \quad i = 1, 2,
\end{align*}
where \( C_{2i} \) is the normalizing constant. From the above identities we have
\begin{align*}
C_{21} \exp\{-OR_1(y)\} P_1(r = 1|y = 0)P_1(y|z, r = 0) = C_{22} \exp\{-OR_2(y)\} P_2(r = 1|y = 0)P_2(y|z, r = 0),
\end{align*}
\begin{align*}
P_1(r = 1) &= P_2(r = 1).
\end{align*}

Note \( \mathbb{E}[\exp\{-OR_i(y)\}] = P_i(y = 0|r = 0)/P_i(y = 0|r = 1) \), we have
\begin{align*}
C_{2i} P_i(r = 1|y = 0) &= P_i(r = 0) \frac{P_i(r = 1|y = 0)}{P_i(r = 0|y = 0)} = \frac{P_i(r = 1)}{\mathbb{E}[\exp\{-OR_i(y)\}|r = 0]}.
\end{align*}
and thus almost surely
\[ \frac{P_1(y|z, r = 0)}{P_2(y|z, r = 0)} = \frac{\mathbb{E}[\exp\{-OR_1(y)\}|r = 0] \exp\{OR_1(y)\}}{\mathbb{E}[\exp\{-OR_2(y)\}|r = 0] \exp\{OR_2(y)\}}. \]

The equation contradicts the condition that the candidates of \( P_{Y|Z,0} \) and \( OR_Y \) make the ratios unequal. So, that unequal ratios is equivalent to the impossibility of two sets of candidates satisfying the same observed quantities, i.e. identification of the joint distribution.

The proof of Theorem 5 relies on the following lemma.

**Lemma 3.** If any element of \( P_{Y|Z,0} \) satisfies:

\[ \int P(y|z, r = 0)h(y)dy = 0 \quad \text{for any } z \quad \Rightarrow \quad h(y) = 0, \]  

then the ratio of any two elements varies with \( z \), i.e. for any \( P_1(y|z, r = 0), P_2(y|z, r = 0) \in P_{Y|Z,0} \), \( P_1(y|z, r = 0)/P_2(y|z, r = 0) \) varies with \( z \).

**Proof.** We prove the lemma by contradiction. Suppose we have two candidates \( P_1(y|z, r = 0) \) and \( P_2(y|z, r = 0) \) such that \( P_1(y|z, r = 0)/P_2(y|z, r = 0) = \tilde{h}(y) \) for some function \( \tilde{h} \). Then we have

\[ \int P_2(y|z, r = 0)dy = \int P_1(y|z, r = 0)dy = \int P_2(y|z, r = 0)\tilde{h}(y)dy = 1, \]

and thus \( \int P_2(y|z, r = 0)\{\tilde{h}(y) - 1\}dy = 0 \). Since \( \tilde{h} - 1 \neq 0 \), the above equation contradicts condition (15). As a result, the ratio of any two elements of \( P_{Y|Z,0} \) must vary with \( z \).

**Proof of Theorem 5**

**Proof.** We prove Theorem 5 under the regularity conditions (1) and (2).

(i) We first assume that \( P(y|z, r = 1) \) follows model (5). We consider \( h \) that

\[ \int P(y|z, r = 0)h(y)dy = 0 \quad \text{for any } z. \]  

From equation (9), we have

\[ P(y|z, r = 0) = \frac{\exp\{OR(y)\}P(y|z, r = 1)P(z|r = 1)}{\mathbb{E}[\exp\{OR(y)\}|r = 1] P(z|r = 0)}. \]
and thus from equation (16), \[
\int P(y|z, r = 1) \exp\{\text{OR}(y)\} h(y) dy = 0 \text{ for any } z. \]
Since \(P(y|z, r = 1)\) follows model (5), by the same proof as that of Theorem 3 we have \(\exp\{\text{OR}(y)\} h(y) = 0\), and thus \(h(y) = 0\). As a result, we have proved that \(\int P(y|z, r = 0) h(y) dy = 0 \text{ for any } z \Rightarrow h = 0\), and by Lemma 3, the ratio of any two elements of \(P_{Y|Z,0}\) varies with \(z\).

(ii) We then assume that \(P(y|z, r = 0)\) follows model (5). By the same proof as that of Theorem 3 we have

\[
\int P(y|z, r = 0) h(y) dy = 0 \text{ for any } z \Rightarrow h = 0,
\]
and thus by Lemma 3, the ratio of any two elements of \(P_{Y|Z,0}\) varies with \(z\).

Proof of Theorem 6

Proof. Since we assume the joint distribution of \((Z, Y, R)\) conditional on \(X\) satisfies the identification condition in Theorem 2, the joint distribution is identifiable. Assuming the estimating equations (11), (12), (13) have a unique solution, we only need to prove that the equations are asymptotically unbiased. We use the subscript “\(\ast\)” to denote the probability limit of the estimators.

Proof of (a). By the law of iterated expectations, we have

\[
\mathbb{E}\left[\{w(X,Y; \alpha, \gamma) R - 1\} h(X,Z)\right] = \mathbb{E}\left[\mathbb{E}\left[\{w(X,Y; \alpha, \gamma) R - 1\} h(X,Z) | X,Y\right]\right].
\]
Assumption 1 implies that

\[
\mathbb{E}\left[\{w(X,Y; \alpha, \gamma) R - 1\} h(X,Z) | X,Y\right] = \mathbb{E}\{w(X,Y; \alpha, \gamma) \mathbb{E}(R|X,Y) - 1\} \mathbb{E}\{h(X,Z) | Y, X\}.
\]

Since \(P(r = 1|y = 0, x; \alpha)\) and \(\text{OR}(y, x; \gamma)\) are correctly specified, under the true values of \(\alpha\) and \(\gamma\), we have \(w(X,Y; \alpha, \gamma) = 1/\mathbb{E}(R|X,Y)\), and thus \(\mathbb{E}\{w(X,Y; \alpha, \gamma) R - 1\} h(X,Z) = 0\), which is the probability limit of equation (11). So equation (11) is asymptotically unbiased.
Proof of (b). By the law of iterated expectations, we have at the true value of $\beta$ and $\gamma$

$$\mathbb{E} \left\{ (1 - R) h(Z, X) | X \right\} = P(r = 0 | X) \mathbb{E} \{ h(Z, X) | r = 0, X \}$$

$$= P(r = 0 | X) \mathbb{E} \{ h(Z, X) | r = 0, X ; \beta, \gamma \}$$

$$= \mathbb{E} \left[ (1 - R) \mathbb{E} \{ h(Z, X) | r = 0, X ; \beta, \gamma \} | X \right],$$

thus,

$$\mathbb{E} \left\{ (1 - R) \{ h(Z, X) - \mathbb{E} \{ h(Z, X) | r = 0, X ; \beta, \gamma \} \} \right\} = 0.$$ 

Therefore equation (12) is asymptotically unbiased and $\hat{\gamma}_{\text{reg}}$ is consistent. As a result, $\mathbb{E} \{ Y | r = 0, x ; \hat{\beta}, \hat{\gamma}_{\text{reg}} \}$ is consistent with $\mathbb{E} \{ Y | r = 0, x \}$, and thus by the law of large numbers, $\hat{\mu}_{\text{reg}}$ converges to

$$\mathbb{E} \left\{ (1 - R) \mathbb{E} \{ Y | r = 0, x \} + R \mathbb{E} \{ Y | r = 1, x \} \right\} = \mathbb{E} \{ Y \},$$

which means $\hat{\mu}_{\text{reg}}$ is consistent.

Proof of (c). Provided $OR(y, x ; \gamma)$ is correctly specified, and $\hat{\beta}$ is consistent, we prove that equations (13) are asymptotically unbiased if either (a) or (b) holds.

(i) If $P(r = 1 | y = 0, x ; \alpha)$ is correctly specified, then $\hat{\alpha}$ is consistent as we have proved that equations (11) are asymptotically unbiased in (a). By the assumption: $Z \perp R | (Y, X)$, we have

$$\mathbb{E} \{ \{ w(X, Y ; \alpha, \gamma) R - 1 \} \{ h(Z, X) - \mathbb{E} \{ h(Z, X) | r = 0, X ; \beta^*, \gamma \} \} \}$$

$$= \mathbb{E} \{ \mathbb{E} \{ w(X, Y ; \alpha, \gamma) R - 1 | X, Y \} \}$$

$$\times \mathbb{E} \{ h(Z, X) - \mathbb{E} \{ h(Z, X) | r = 0, X ; \beta^*, \gamma \} | X, Y \},$$

which equals 0 since $\mathbb{E} \{ w(X, Y ; \alpha, \gamma) R - 1 | X, Y \} = 0$.

(ii) If $P(z, y | r = 1, x ; \beta)$ is correctly specified, and $\hat{\beta}$ is consistent, we have $\mathbb{E} \{ h(Z, X) | r = 0, X \} = \mathbb{E} \{ h(Z, X) | r = 0, X ; \beta, \gamma \}$. Note that

$$\mathbb{E} \{ \{ w(X, Y ; \alpha^*, \gamma) R - 1 \} \{ h(Z, X) - \mathbb{E} \{ h(Z, X) | r = 0, X ; \beta, \gamma \} \} \}$$

$$= \mathbb{E} \{ \left[ R \{ w(X, Y ; \alpha^*, \gamma) - 1 \} \{ h(Z, X) - \mathbb{E} \{ h(Z, X) | r = 0, X ; \beta, \gamma \} \} \right] \}$$

$$- \mathbb{E} \{ (1 - R) \{ h(Z, X) - \mathbb{E} \{ h(Z, X) | r = 0, X ; \beta, \gamma \} \} \}. $$

The second term of the right hand side equals 0 as we have proved in (b). We only need to prove that the first term also equals 0. From equation (3) we have

$$R \{ w(X, Y ; \alpha^*, \gamma) - 1 \} = R \exp \{ OR(Y | X ; \gamma) \} \frac{P(r = 0 | y = 0, X ; \alpha^*)}{P(r = 1 | y = 0, X ; \alpha^*)}. $$
and note equation (9)

\[ P(z, y| r = 0, x) = \frac{\exp\{OR(y|x)\}P(z, y| r = 1, x)}{E[\exp\{OR(Y|x)\}|r = 1, x]}, \]

we have

\[
\begin{align*}
\mathbb{E}\{h(Z, x)|r = 0, x\} &= \int \int h(z, x)P(z, y| r = 0, x)dydz \\
&= \frac{\int \exp\{OR(y|x)\}P(z, y| r = 1, x)h(z, x)dydz}{E[\exp\{OR(Y|x)\}|r = 1, x]} \\
&= \frac{E[\exp\{OR(Y|x)\}h(Z, x)|r = 1, x]}{E[\exp\{OR(Y|x)\}|r = 1, x]} \\
&= \frac{E[R\exp\{OR(Y|x)\}h(Z, x)|x]}{E[R\exp\{OR(Y|x)\}|x]}.
\end{align*}
\]

So we have

\[
\mathbb{E}[R\exp\{OR(Y|x)\}h(Z, x)|x] = \mathbb{E}[R\exp\{OR(Y|x)\}|x]\mathbb{E}\{h(Z, x)|r = 0, x\},
\]

and thus

\[
\mathbb{E}[R\exp\{OR(Y|x)\}\{h(Z, x) - \mathbb{E}[h(Z, x)|r = 0, x]\}|x] = 0,
\]

\[
\mathbb{E}\left[ \frac{P(r = 0|y = 0, X; \alpha^*)}{P(r = 1|y = 0, X; \alpha^*)} \cdot R\exp\{OR(Y|X; \gamma)\} \cdot \{h(Z, X) - \mathbb{E}[h(Z, X)|r = 0, X; \beta, \gamma]\} \right] = 0.
\]

By the law of iterated expectations, we have at the truth of \(\beta\) and \(\gamma\)

\[
\mathbb{E}\left[ \frac{P(r = 0|y = 0, X; \alpha^*)}{P(r = 1|y = 0, X; \alpha^*)} \cdot R\exp\{OR(Y|X; \gamma)\} \cdot \{h(Z, X) - \mathbb{E}[h(Z, X)|r = 0, X; \beta, \gamma]\} \right] = 0.
\]

So the first term must equal 0. Therefore, \(\hat{\gamma}_{dr}\) is doubly robust. By similar arguments as (i) and (ii) with \(h(Z, X)\) replaced by \(Y\), we can prove that under the conditions of (a)

\[
(17) \quad \mathbb{E}[\{w(X, Y; \alpha, \gamma)R - 1\} \{Y - \mathbb{E}[Y|r = 0, X; \beta^*, \gamma]\}] = 0,
\]
and thus
\[ \mu = E\{w(X,Y; \alpha, \gamma) R \{Y - E[Y|r = 0, X; \beta^*, \gamma]\}\} + E\{Y|r = 0, X; \beta, \gamma]\}; \]
and under the conditions of (b)
\[ (18) \quad E\{w(X,Y; \alpha^*, \gamma) R - 1\}\{Y - E[Y|r = 0, X; \beta, \gamma]\}\} = 0, \]
and thus
\[ \mu = E\{w(X,Y; \alpha^*, \gamma) R \{Y - E[Y|r = 0, X; \beta, \gamma]\}\} + E\{Y|r = 0, X; \beta, \gamma]\}. \]
Therefore, \( \hat{\mu}_{dr} \) is doubly robust for the mean of the outcome.

SUPPLEMENTARY MATERIAL

Supplement to “Identification and Doubly Robust Estimation of Data Missing Not at Random With a Shadow Variable”
(doi: COMPLETED BY THE TYPESETTER; .pdf). This supplement contains variance estimation and additional details on inference, details of Example 1, 2, simulation results for a binary outcome, and details on model fits for the real data example.

REFERENCES

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–973.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press Baltimore.

Das, M., Newey, W. K. and Vella, F. (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies* **70** 33–58.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 1–38.

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

Fay, R. E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association* **81** 354–365.

Greenlees, J. S., Reece, W. S. and Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association* **77** 251–261.

Hall, A. R. (2005). *Generalized Method of Moments*. Oxford University Press Oxford.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47** 153–161.
HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**(159) 663–685.

HU, Y. and SHIU, J.-L. (2011). Nonparametric identification using instrumental variables: sufficient conditions for completeness. *Department of Economics Working Paper, 581, Johns Hopkins University*.

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.

LITTLE, R. J. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88**(421) 125–134.

LITTLE, R. J. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika* **81**(3) 471–483.

LITTLE, R. J. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley: New York.

MA, W. Q., GENG, Z. and HU, Y. H. (2003). Identification of graphical models for nonignorable nonresponse of binary outcomes in longitudinal studies. *Journal of multivariate analysis* **87**(2) 24–45.

MIAO, W., DING, P. and GENG, Z. (2014). Identifiability of normal and normal mixture models with nonignorable missing data. Submitted.

NAUGHTON, B. (2007). *The Chinese Economy: Transitions and Growth*. MIT press.

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**(426) 846–866.

ROBINS, J. M., ROTNITZKY, A. and SCHARFSTEIN, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials* 1–94. Springer.

ROTNITZKY, A., ROBINS, J. M. and SCHARFSTEIN, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93**(444) 1321–1339.

RUBIN, D. B. (1976). Inference and missing data (with discussion). *Biometrika* **63**(3) 581–592.

RUBIN, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys* **81**. John Wiley & Sons.

SCHARFSTEIN, D. O. and IRIZARRY, R. A. (2003). Generalized additive selection models for the analysis of studies with potentially nonignorable missing outcome data. *Biometrics* **59**(3) 601–613.

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94**(446) 1096–1120.

SCHENKER, N. and WELSH, A. (1988). Asymptotic results for multiple imputation. *The Annals of Statistics* 1550–1566.

TCHETGEN TCHETGEN, E. J. and WIRTH, K. (2013). A general instrumental variable framework for regression analysis with outcome missing not at random.

TSIATIS, A. (2007). *Semiparametric Theory and Missing Data*. Springer Science & Business Media.

VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer Science & Business Media.

VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. Cambridge university press.
Vansteelandt, S., Rotnitzky, A. and Robins, J. (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika* **94** 841–860.

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.

Zahner, G. E., Pawelkiewicz, W., DeFrancesco, J. J. and Adnpoz, J. (1992). Children’s mental health service needs and utilization patterns in an urban community: an epidemiological assessment. *Journal of the American Academy of Child & Adolescent Psychiatry* **31** 951–960.

Zhao, J. and Shao, J. (2014). Semiparametric pseudo likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association* accepted.

Wang Miao
Beijing International Center for Mathematical Research
Beijing University
Haidian District, Beijing 100871
E-mail: mwfy@pku.edu.cn

Eric Tchetgen Tchetgen
Department of Biostatistics
Harvard University
Boston, Massachusetts 02115
E-mail: etchetge@hsph.harvard.edu

Zhi Geng
School of Mathematical Sciences
Center for Statistical Science
Beijing University
Haidian District, Beijing 100871
E-mail: zhigeng@pku.edu.cn