Statistica Sinica Preprint No: SS-2016-0325.R2	
Title	Discrete Choice Models for Nonmonotone Nonignorable Missing Data: Identification and Inference
Manuscript ID	SS-2016-0325.R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0325
Complete List of Authors	Eric J. Tchetgen Tchetgen Linbo Wang and BaoLuo Sun
Corresponding Author	Eric J. Tchetgen Tchetgen
E-mail	etchetgen@gmail.com

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0325

Original Article

Discrete Choice Models for Nonmonotone Nonignorable Missing Data: Identification and Inference

Eric J. Tchetgen Tchetgen, Linbo Wang, BaoLuo Sun

Department of Biostatistics,

Harvard University

Abstract

Nonmonotone missing data arise routinely in empirical studies of the social and health sciences and, when ignored, can induce selection bias and loss of efficiency. It is common to account for nonresponse under a missing-at-random assumption which, although convenient, is rarely appropriate when nonresponse is nonmonotone. Likelihood and Bayesian missing data methodologies often require specification of a parametric model for the full data law, thus *a priori* ruling out any prospect for semiparametric inference. In this paper, we propose an all-purpose approach which delivers semiparametric inferences when missing data are nonmonotone and not at random. The approach is based on a discrete choice model (DCM) as a means to generate a large class of nonmonotone nonresponse mechanisms that are nonignorable. Sufficient conditions for nonparametric identification are given, and a general framework for fully parametric and semiparametric inference under an arbitrary DCM is proposed. Special consideration is given to the case of logit discrete choice nonresponse model (LDCM) for which we describe generalizations of inverse-probability weighting, pattern-mixture estimation, doubly robust estimation, and multiply robust estimation.

KEY WORDS: missing not at random, nonmonotone missing data, pattern mixture, doubly robust, inverse-probability-weighting.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0325

1 Introduction

Missing data are a common occurence in empirical research in the health and social sciences, and often affect one's ability to draw reliable inferences whether from an experimental or nonexperimental study. Non-response can occur in sample surveys, due to dropout or non-compliance in clinical trials, or due to data excision by error or in order to protect confidentiality. In many situations, there may be no nested pattern of missingness such that observing variable X_k implies that variable X_j is also observed, for any $j < k$. Nonmonotone missing data patterns may occur, for instance, when individuals who dropped out of a longitudinal study re-enter at later time points; likewise, in regression analysis nonmonotone nonresponse may occur if the outcome or any of the regressors may be unobserved for a subset of the sample in an arbitrary pattern. Missing data are said to be completely-at-random (MCAR) if the nonresponse process is independent of both observed and unobserved variables in the full data, and missing-at-random (MAR) if, conditional on observed variables under a nonresponse pattern, the probability of observing the pattern does not depend on unobserved variables under the pattern (Rubin (1976); Little and Rubin (2002), Robins et al. (1994)). A nonresponse process which is neither MCAR nor MAR is said to be missing-not-at-random (MNAR).

While complete-case analysis is perhaps the most widely-used method to handle missing data, the approach is generally not recommended as it can give biased inferences when nonresponse is not MCAR. Formal methods to appropriately account for incomplete data include fully parametric likelihood and Bayesian approaches (Little and Rubin (2002); Horton and Laird (1999); Ibrahim and Chen (2000); Ibrahim et al. (2002, 2005)) which are most commonly implemented under MAR using the EM algorithm or via multiple imputation (MI) (Dempster et al. (1977), Rubin (1977); Schafer (1997)). Inverse probability weighting (IPW) is another approach to accounting for selection bias due to missing data (Horvitz and Thompson (1952); Robins et al. (1994); Tsiatis (2006)). While IPW estimation avoids specification of a full-data likelihood, the approach does require a model for the nonresponse process. However, the development of general coherent models for nonmonotone nonresponse has proved to be particularly challenging, even under the MAR assumption; see Robins and Gill (1997) and Sun and Tchetgen Tchetgen (2016) for two

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0325

concrete proposals and further discussion.

Despite recent progress in development of MAR methodology, as argued by Gill and Robins (1997), Robins (1997) and Little and Rubin (2002), the assumption is generally hard to justify on substantive grounds when nonresponse is nonmonotone. Instead, allowing for MNAR data seems particularly befitting in the context of nonmonotone nonresponse and has received substantial attention, particularly in the context of fully parametric models (Deltour et al. (1999), Albert (2000), Ibrahim et al. (2001), Fairclough et al. (1998), Troxel et al. (1998), Troxel, Lipsitz and Harrington (1998)). MNAR approaches which do not necessarily rely on parametric assumptions have also been developed in recent years. Notable examples include the group permutation model (GPM) of Robins (1997) and the block conditional MAR (BCMAR) model of Zhou et al. (2010). These approaches allow for non-ignorable missing data in the sense that the nonresponse process of a given variable may depend on values of other missing variables. However, neither BCMAR nor GPM allows the missingness probability of a given variable to depend on the value of the variable. Based on subject matter considerations, it is often desirable to consider non-ignorable processes where the missingness probability of a variable depends on the possibly unobserved value of the variable, therefore, methods for non-ignorable missing data mechanisms beyond BCMAR and GPM are of interest.

In this paper, we propose a large class of non-ignorable nonmonotone nonresponse models that unlike BCMAR and GPM, do not *a priori* rule out the possibility that the probability of observing a given variable depends on the unobserved value of the variable. Our approach is based on so-called discrete choice models (DCM). DCMs were first introduced and predominantly used, in economics and other social sciences, as a principled approach for generating a large class of multinomial models to describe discrete choice decision making under rational utility maximization. Here DCMs are used as a means to generate a large class of nonmonotone nonresponse mechanisms which are nonignorable. Sufficient conditions for nonparametric identification are given, and a general framework for semiparametric inference under an arbitrary DCM is proposed. Special consideration is given to the case of logit discrete choice nonresponse model (LDCM). Our identification condition in the case of the LDCM, states that the conditional distribution of unobserved variables given

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0325

observed variables for any nonresponse pattern, matches the corresponding conditional distribution in complete-cases. This assumption is equivalent to the well-known complete-case missing value (CCMV) restriction in the pattern mixture (PM) literature that has previously been developed for fully likelihood-based inference (Little (1993)). Therefore, our approach provides a comprehensive treatment of semiparametric inference for MNAR nonresponse under Little's CCMV restriction. In addition to reviewing Little's PM likelihood approach, we describe a generalization of inverse-probability weighting (IPW), and both doubly robust (DR) and multiply robust (MR) estimation, that are the nonmonotone MNAR analogues of existing results for monotone MAR nonresponse (Tsiatis (2006)). We establish that whenever J nonresponse patterns are observed, the proposed LDCM DR estimators can be made 2^J-robust in the sense that, for each nonresponse pattern, valid inferences can be obtained if one of two pattern-specific models is correctly specified but not necessarily both. As far as we know, our paper represents the first instance of a doubly (2^J-) robust estimator obtained for a general nonmonotone nonignorable missing data model that is identified from the observed data alone. Our proposed inferences under the LDCM are quite attractive as a generic nonignorable approach for arbitrary nonmonotone patterns, mainly because they are somewhat easy to implement, have good robustness properties, and appear to have good finite sample performance as we illustrate via simulation studies and an HIV data application. In closing, we briefly consider IPW inference for DCMs outside of the LDCM, which can generally be used to account for nonmonotone nonignorable missing data even when Little's CCMV condition fails and therefore the LDCM may not be appropriate.

2 Notation and definitions

Suppose full data consist of n i.i.d. realizations of a random K-vector $L = (L_1, ..., L_K)'$. Let R denote the scalar random variable encoding missing data patterns, and J denote the total number of observed patterns. For missing data pattern $R = r$, where $1 \leq r \leq J \leq 2^K$, we use $L_{(r)}$ and $L_{(-r)}$ to denote observed and unobserved components of L, respectively so that $L = (L_{(r)}, L_{(-r)})$. We reserve $r = 1$ to denote complete cases. Throughout, write $\Pr\{R = r|L\} = \pi_r(L) = \Pi_r$ for all r. For each realization, we observe $\left(R, L_{(R)}\right)$. For instance, if the full data L is a bivariate

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0325

binary vector (L_1, L_2) and $J = 3$ nonmonotone nonresponse patterns are observed in the sample: $R = 1$,$L_{(1)} = L$; $R = 2$, $L_{(2)} = L_1$; and $R = 3, L_{(3)} = L_2$.

Throughout, we make the positivity assumption,

$$\Pi_1 > \sigma > 0 \text{ a.s.}, \tag{1}$$

for a fixed positive constant σ. Assumption (1) is needed for nonparametric identification of the full data distribution, and its smooth functionals as well as finite asymptotic variance of IPW estimators (Robins et al. (1999)). As discussed in Section 2.3, complete-case IPW relies on obtaining a consistent estimator of $\pi_1(L) = 1 - \sum_{r \neq 1} \pi_r(L)$ which in turn requires estimating the nonresponse process $\{\pi_r(L) : r\}$. The nonresponse process clearly fails to be nonparametrically identified under (1) alone. In the next section, we describe a set of sufficient conditions to identify a model for the complete-case probability $\pi_1(L)$ under the discrete choice framework when missingness is nonmonotone and not at random.

Our first result provides a generic nonparametric representation of the joint law of $f(R, L)$ that will be used throughout. The result adapts the generalized odds ratio parametrization of a joint distribution due to Chen (2010) to the missing data context; see also Tchetgen Tchetgen et al. (2010). Let $\text{Odds}_r(L) = \pi_r(L)/\pi_1(L)$.

Lemma 1 *We have that*

$$f(R, L) = \frac{\prod_{r \neq 1} \text{Odds}_r(L)^{I(R=r)} f(L|R = 1)}{\iint \prod_{r \neq 1} \text{Odds}_r(l^*)^{I(r^*=r)} f(l^*|R = 1) \, d\mu(r^*, l^*)},$$

provided $\iint \prod_{r \neq 1} \text{Odds}_r(l^*)^{I(r^*=r)} f(l^*|R = 1) \, d\mu(r^*, l^*) < \infty$, *with* μ *a dominating measure of the CDF of* (R, L).

Lemma 1 clarifies what the identification task entails: under (1), $f(L|R = 1)$ is just-identified, and therefore $f(R, L)$ is nonparametrically just-identified only if one can just-identify $\text{Odds}_r(L)$ for all r. Below we describe a sufficient condition for identification under the discrete choice model

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0325

of the nonresponse process .

3 Identification

3.1 The discrete choice nonresponse model

The DCM associates with each realized nonresponse pattern $r = 1, ..., J \leq 2^K$ an underlying utility function $U_r = \mu_r(L) + \varepsilon_r$, where $\{\varepsilon_r : r\}$ are i.i.d. with cumulative distribution function F_ε, and $\mu_r(L)$ encodes the dependence of a person's utility on L (McFadden (1984), Train (2009)). Some common choices of F_ε include the extreme value distribution (further discussed below) and the normal distribution, although in principle any CDF could be specified. It is then assumed that a person's observed response pattern maximizes her utility, $R = \arg\max_r \{U_r : r\}$. Together, these assumptions imply that for each r,

$$\Pi_r = \pi_r(L) = \Pr(R = r|L) = \int \prod_{s \neq r} F_\varepsilon(\Delta\mu_{rs}(L) + \varepsilon)\, dF_\varepsilon(\varepsilon), \tag{2}$$

where $\Delta\mu_{rs}(L) = \mu_r(L) - \mu_s(L)$ captures the dependence on L of a difference in utility in comparing a person's choice between nonresponse patterns r and s, see Train (2009). The integral in (2) is generally not available in closed form for most choices of F_ε (with the notable exception of the extreme value distribution, see Section 2.2), but can easily be evaluated by numerical integration using, say, Gaussian quadrature. Although not immediately apparent from the expression in the display, (2) gives rise to a proper probability mass function, that is $\sum_r \pi_r(l) = 1$ for all values of l and for any choice of F_ε. This remarkable result is a direct consequence of utility maximization as a formal principle for generating multinomial probabilities $\{\pi_r : r\}$. Another observation is that only differences in utility matter in determining the choice probabilities; in other words, the absolute level of a person's utility for a given nonresponse pattern is irrelevant and only relative utility drives the choice of a nonresponse pattern over another. Clearly, model (2) is not identifiable without an additional assumption, even given knowledge of F_ε.

For the purpose of identification, we consider the assumption that the relative utility $\Delta\mu_{1r}(L)$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0325

of any nonresponse pattern $r \neq 1$ compared with that of complete-case pattern $r = 1$, only depends on data observed under both patterns,

$$\Delta\mu_{1r}(L) = \Delta\mu_{1r}\left(L_{(r)}\right) \text{ for all } r \text{ almost surely.} \tag{3}$$

The assumption essentially states that when faced with the choice between nonresponse pattern $r \neq 1$ versus providing complete data, the excess utility a subject would experience choosing one over the other only depends on data observed under both choices. Under this assumption, one may write

$$\Pi_r = \int \prod_{s \neq r} F_\varepsilon\left(\Delta\mu_{1s}\left(L_{(s)}\right) - \Delta\mu_{1r}\left(L_{(r)}\right) + \varepsilon\right) dF_\varepsilon(\varepsilon) \tag{4}$$

Even under (3), Π_r generally depends on unobserved variables for all r, and therefore, data are missing not at random, and the corresponding observed data likelihood is nonignorable. Nevertheless, as we show in Section 5, given any continuous F_ε, equation (4) is nonparametrically identified for each r provided (1) holds. We leave the detailed discussion of inference under user-specified F_ε to Section 5, instead, to fix ideas, we further discuss identification and inference under the logit DCM.

3.2 The logit discrete choice model

When F_ε is the extreme value distribution, the integral in (2) is available in closed-form, and gives the following logit DCM (Train (2009)): $\pi_r(L) = \text{Odds}_r(L)/(1 + \sum_{s \neq 1} \text{Odds}_s(L))$, where $\text{Odds}_r(L) = \exp(\Delta\mu_{1r}(L))$ for all r. Under (3), $\text{Odds}_r(L) = \text{Odds}_r\left(L_{(r)}\right)$, and therefore

$$\Pi_r = \frac{\text{Odds}_r\left(L_{(r)}\right)}{1 + \sum_{s \neq 1} \text{Odds}_s\left(L_{(s)}\right)}, \text{ for all } r \neq 1. \tag{5}$$

To illustrate (5), consider an example with $L = (L_1, L_2, L_3)$. Suppose that there are four nonresponse patterns, $L_{(1)} = L, L_{(2)} = (L_1, L_2), L_{(3)} = L_3, L_{(4)} = \varnothing$. Then, by (3) $\text{Odds}_2(L) = \text{Odds}_2\left(L_{(2)}\right); \text{Odds}_3(L) = \text{Odds}_3\left(L_{(3)}\right); \text{Odds}_4(L) = \text{Odds}_4\left(L_{(4)}\right) = \text{Odds}_4$ is a constant. Furthermore, according to (5) $\Pi_2 = \text{Odds}_2\left(L_{(2)}\right)/c(L); \Pi_3 = \text{Odds}_3\left(L_{(3)}\right)/c(L)$; $\Pi_4 = \text{Odds}_4/c(L)$,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0325

where $c(L) = \left(1 + \sum_{s \neq 1} \text{Odds}_s\left(L_{(s)}\right)\right)$. Therefore, by virtue of $c(L)$, the nonresponse probabilities Π_j, $j = 2, 3, 4$ are each a function of $\tilde{L} = \cup_{j=2,3,4} L_{(j)}$, the union set of observed variables across all the nonresponse patterns. Since the variable set $\tilde{L} \setminus L_{(j)}$ is not observed for each of the missing data patterns $j = 2, 3, 4$, the nonresponse process is clearly MNAR. In particular, Π_4 is a function of $\tilde{L}$ even though no variable is observed in the fourth missing data pattern.

An equivalent characterization of (5) is

$$L_{(-r)}|R = r, L_{(r)} \sim L_{(-r)}|R = 1, L_{(r)} \quad \text{for all } r \neq 1, \tag{6}$$

the conditional distribution of unobserved variables $L_{(-r)}$ given observed variables $L_{(r)}$ for nonresponse pattern r matches the corresponding conditional distribution among complete-cases. Although the LDCM is derived as a particular DCM, one could in principle take (6) as primitive identifying condition without necessarily making reference to a DCM and the existence of its associated variables $\{\varepsilon_r : r\}$ .This amounts to nonparametric identification under the complete-case missing value restriction of Little (1993). As shown in Section 5, adoption of the more general DCM framework is advantageous as it gives rise to a richer class of nonresponse models and facilitates identification; in fact, a different choice for the distribution F_ε corresponds to a nonmonotone not at random nonresponse model which does not generally satisfy Little's CCMV restriction but is nevertheless just-identified under (1) and (3).

It is instructive to compare (6) to standard MAR, which states that

$$L_{(-r)}|R = r, L_{(r)} \sim L_{(-r)}|L_{(r)} \quad \text{for all } r \neq 1. \tag{7}$$

Clearly, (6) and (7) have fundamentally different implications for inference. Specifically, when the nonresponse process and the full data distribution depend on separate parameters, the MAR assumption implies that the part of the observed data likelihood which depends on the full data parameter factorizes from the nonresponse process. The missing data mechanism is then said to be "ignorable" (Little and Rubin (2002)) because it is possible to learn about the full data law without necessarily estimating the missing data process or, equivalently, it is possible to learn

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0325

about the missing data process without modeling the full data law (Sun and Tchetgen Tchetgen (2016)). No such factorization is in general available under CCMV as the missing data process is nonignorable. In spite of possible challenges due to lack of factorization, we show that estimation of nonmonotone non-response mechanisms under (6) is nevertheless relatively straightforward. Furthermore, assumption (6) is invariant to the number and nature of other nonresponse patterns potentially realized in the observed data. In contrast, MAR does not enjoy a similar invariance property because addition or deletion of a nonresponse pattern from the observed sample changes the interpretation of (7) as it implies marginalizing over a different set of nonresponse patterns to obtain its right-hand side. Finally, (6) and (7) only coincide when there is a single nonresponse pattern, $J = 2$.

Remark 1 Sun and Tchetgen Tchetgen (2016) proposed an approach tailored specifically to model a nonmonotone nonresponse process under (7), but did not consider (3). As (3) and (7) differ, their approach cannot be used under (3).

Lemma 2 *If assumptions* (1) *and* (2) *hold with* F_ε *the extreme value distribution, and if* (3) *holds, the joint distribution* $f(R, L)$ *is nonparametrically just-identified from the observed data* (L_R, R), *with*

$$f(R,L) = \frac{\prod_{r\neq 1} \text{Odds}_r\left(L_{(r)}\right)^{I(R=r)} f\left(L|R=1\right)}{\iint \prod_{r\neq 1} \text{Odds}_r\left(l^*_{(r)}\right)^{I(r^*=r)} f\left(l^*|R=1\right) d\mu\left(r^*, l^*\right)}, \tag{8}$$

where μ *is a dominating measure of the CDF of* (R, L).

This explicit expression for $f(R, L)$ appears to be new, and can be used to compute the full data density $f(L) = \sum_r f(r, L)$. Equation (8) can also be used for maximum likelihood estimation. Specifically, let $f\left(L|R=1;\eta\right)$ denote a parametric model for $f\left(L|R=1\right)$ with unknown parameter η, and consider a parametric model for nonresponse process $\Pi_r\left(\alpha\right) = \text{Odds}_r\left(L_{(r)};\alpha_r\right)/\{1 + \sum_{s\neq 1}\text{Odds}_s\left(L_{(s)};\alpha_s\right)\}$ with unknown parameter $\alpha = \{\alpha_r : r\}$, where α_r indexes a parametric model for $\text{Odds}_r\left(L_{(r)};\alpha_r\right)$. Let $f(R, L;\theta)$ denote the corresponding model for $f(R, L)$, where $\theta = (\eta, \alpha)$. The maximum likelihood estimator (MLE) $\hat{\theta}_{mle}$ maximizes the observed data log-likelihood $\mathbb{P}_n \log \int f(R, L;\theta)d$ where $\mathbb{P}_n\left(\cdot\right) = n^{-1}\sum_i\left(\cdot\right)_i$. The full data likelihood $f(L;\hat{\theta}_{mle}) = \int f(r, L;\hat{\theta}_{mle})d\mu\left(r\right)$ can then be

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0325

used to make inferences about a given full data functional of interest according to the plug-in principle. By standard likelihood theory, the MLE is asymptotically efficient in the model $\mathcal{M}_{lik}$ corresponding to the set of laws $\{f(R, L; \theta) : \theta\}$. A major drawback of maximum likelihood inference is its lack of robustness to model mis-specification, because $\hat{\theta}_{mle}$ is likely inconsistent if either $\Pi_r(\alpha)$ or $f(L|R = 1; \eta)$ is incorrectly specified. Below, we consider four semiparametric estimators which are potentially more robust than direct likelihood maximization.

4 Semiparametric Inference

4.1 Inverse-probability weighting estimation

Suppose the parameter of interest, β_0, is the unique solution to the full data population estimating equation $E\{U(L; \beta_0)\} = 0$, where the expectation is taken over the distribution of the complete data L. In principle, no further restriction on the distribution of L is strictly required; in fact, estimation is possible under certain weak regularity conditions (van der Vaart (1998)) as long as a full data unbiased estimating function exists. In the presence of missing data, the estimating function can only be evaluated for complete-cases that might be highly selected even under MAR. This motivates the use of IPW estimating functions of complete-cases to form the complete-case population estimating equation

$$E\left\{\frac{1(R=1)}{\Pi_1}U(L;\beta_0)\right\} = 0, \tag{9}$$

which holds by straightforward iterated expectations. The IPW estimator $\widehat{\beta}_{ipw}$ that solves the empirical version of this equation will in general be inefficient especially when the fraction of complete-cases is relatively small, since incomplete cases are discarded (except when estimating Π_1). In the next section we will describe a strategy to recover information from incomplete-cases by augmenting the estimating function shown in (9) to gain efficiency and, potentially, robustness. The IPW estimating equations framework encompasses a large variety of settings under which investigators may wish to account for non-monotone missing data. These include IPW of the full data score

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0325

equation, where the score function is such an unbiased estimating function, given a model $f(L;\beta_0)$ for the law of the full data, in which case (9) reduces to $E\left\{1\left(R=1\right)\partial\log f(L;\beta)/\partial\beta|_{\beta_0}/\Pi_1\right\}=0$

We now describe a straightforward approach to obtain a consistent estimator of Π_1 in the semiparametric model which specifies a parametric LCDM $\{\Pi_r(\alpha):r\}$, but allows $f(L|R=1)$ to remain unrestricted. We denote this model by $\mathcal{M}_R$. The approach follows as (5) implies that

$$\Pr\left(R=r|L,R\in\{1,r\}\right)=\Pi_{r,c}=\frac{\text{Odds}_r\left(L_{(r)}\right)}{1+\text{Odds}_r\left(L_{(r)}\right)},\text{ for all }r.$$

This also gives the equivalent representation of the CCMV restriction

$$R\perp\!\!\!\perp L_{(-r)}|R\in\{r,1\},L_{(r)}\text{ for each }r.$$

Here $L_{(r)}$ is fully observed for observations $R\in\{1,r\}$. Thus, to estimate the parametric model $\{\Pi_{r,c}(\alpha):r\}$, for each nonresponse pattern r one can fit the logistic regression $\Pi_{r,c}(\alpha_r)=\text{Odds}_r\left(L_{(r)};\alpha_r\right)/\{1+\text{Odds}_r\left(L_{(r)};\alpha_r\right)\}$ by maximum likelihood estimation, restricted to the subset of data containing complete-cases and incomplete-cases of pattern r only. We define the restricted MLE

$$\begin{aligned}\widetilde{\alpha}_r&=\arg\max_{\alpha_r}\mathbb{P}_n\text{llik}_{r,c}(\alpha_r)\\&=\arg\max_{\alpha_r}\mathbb{P}_n\left\{I\left(R=r\right)\log\Pi_{r,c}(\alpha_r)+I\left(R=1\right)\log\left(1-\Pi_{r,c}(\alpha_r)\right)\right\}.\end{aligned}$$

Under assumption (1), the restricted MLE $\widetilde{\alpha}$ is consistent and asymptotically normal under model $\mathcal{M}_R$. The resulting estimator of the complete-case probability Π_1 under $\mathcal{M}_R$ is

$$\Pi_1(\widetilde{\alpha})=\frac{1}{1+\sum_{s\neq1}\text{Odds}_s\left(L_{(s)};\widetilde{\alpha}_s\right)},$$

which, in turn, provides the IPW estimator $\widehat{\beta}_{ipw}$ of β that solves

$$\mathbb{P}_n\left\{U_{ipw}(L_{(R)},R;\widehat{\beta}_{ipw},\widetilde{\alpha})\right\}=0,\tag{10}$$

where $U_{ipw}(L_{(R)},R;\widehat{\beta}_{ipw},\widetilde{\alpha})=1\left(R=1\right)U(L;\widehat{\beta}_{ipw})/\Pi_1(\widetilde{\alpha})$. Under standard regularity conditions,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0325

one can show that, under $\mathcal{M}_R$, the IPW estimator $\widehat{\beta}_{ipw}$ is, in large samples, approximately normal with mean β_0 and asymptotic variance $\hat{\Gamma}_{ipw}^{-1}\hat{\Omega}_{ipw}\hat{\Gamma}_{ipw}^{-1}$, where

$$\hat{\Gamma}_{ipw}^{-1} = -\frac{\partial}{\partial\beta^T}\mathbb{P}_n\left\{U_{ipw}(L_{(R)}, R; \beta, \widetilde{\alpha})\right\}\Big|_{\widehat{\beta}_{ipw}};$$

$$\hat{\Omega}_{ipw} = n^{-1}\mathbb{P}_n\left\{\left[U_{ipw}(L_{(R)}, R; \widehat{\beta}_{ipw}, \widetilde{\alpha}) + \frac{\partial}{\partial\alpha^T}\mathbb{P}_n\left\{U_{ipw}(L_{(R)}, R; \widehat{\beta}_{ipw}, \alpha)\right\}\Big|_{\widetilde{\alpha}}\widehat{IF}_\alpha\right]^{\otimes 2}\right\};$$

$$\widehat{IF}_\alpha = -\left[\frac{\partial^2}{\partial\alpha\partial\alpha^T}\mathbb{P}_n\left\{\sum_{r\neq 1}\text{llik}_{r,c}(\alpha_r)\right\}\Big|_{\widetilde{\alpha}}\right]^{-1}\frac{\partial}{\partial\alpha}\left\{\sum_{r\neq 1}\text{llik}_{r,c}(\alpha_r)\right\}\Big|_{\widetilde{\alpha}}.$$

For inference about a component of β_0, one can report the corresponding Wald-type 95% confidence interval.

4.2 Pattern-mixture LDCM estimation

In this section, we consider an alternative approach for obtaining inferences about the full data parameter β_0 defined in the previous one. The approach is a slight generalization of the pattern-mixture approach due to Little (1993). To proceed, note that

$$\begin{aligned} E\{U(L;\beta_0)\} &= E\left[E\left\{U(L;\beta_0)|R, L_{(R)}\right\}\right], \\ &= E\left[E\left\{U(L;\beta_0)|R=1, L_{(R)}\right\}\right] \\ &= E\left[\sum_r I(R=r)E\left\{U(L;\beta_0)|R=1, L_{(r)}\right\}\right] \\ &= 0, \end{aligned} \tag{11}$$

where the second equality follows from (6) . Now, consider the semiparametric model $\mathcal{M}_L$ that posits a parametric model $f(L|R=1;\eta)$ while allowing the nonresponse process $\{\Pi_r : r\}$ to remain unrestricted. Let $\widetilde{\eta}$ denote the restricted MLE of η in $\mathcal{M}_L$ obtained using only complete-case data, $\widetilde{\eta} = \arg\max_\eta \mathbb{P}_n \text{llik}_{l,c}(\eta) = \arg\max_\eta \mathbb{P}_n I(R=1)\log f(L|R=1;\eta)$. An empirical version of (11) can

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0325

then be used to obtain a pattern mixture estimator $\widehat{\beta}_{pm}$ of β_0,

$$0 = \mathbb{P}_n \left[U_{pm}(L_{(R)}, R; \widehat{\beta}_{pm}, \widetilde{\eta}) \right], \tag{12}$$

where

$$U_{pm}(L_{(R)}, R; \widehat{\beta}_{pm}, \widetilde{\eta}) = \sum_r I(R=r) E\left\{ U(L; \widehat{\beta}_{pm}) | R=1, L_{(r)}; \widetilde{\eta} \right\}, \tag{13}$$

and $E\left\{ U(L; \widehat{\beta}_{pm}) | R=1, L_{(r)}; \widetilde{\eta} \right\} = \int U(l_{(-r)}, L_{(r)}; \widehat{\beta}_{pm}) f\left(l_{(-r)} | L_{(r)} | R=1; \widetilde{\eta} \right) d\mu\left(l_{(-r)} \right)$. To ensure that models $\left\{ f\left(l_{(-r)} | L_{(r)} | R=1; \widetilde{\eta} \right), r \neq 1 \right\}$ are compatible, one may need to specify a model for $f\left(L | R=1 \right)$; this is effectively the approach followed by Little (1993). In the pattern mixture approach, the model for $f(L)$, which is of primary scientific interest, is indirectly specified via models for the various conditional densities $\left\{ f\left(l_{(-r)} | L_{(r)} | R=1 \right), r \neq 1 \right\}$ and the marginal densities $\left\{ f\left(L_{(r)} | R=r \right), r \neq 1 \right\}$ according to the mixture $f(L) = \sum_r f\left(l_{(-r)} | L_{(r)} | R=1 \right) f\left(l_{(r)} | R=r \right) \Pr(R = r)$ (Little (1993)). Under standard regularity conditions, one can show that, in large samples, $\widehat{\beta}_{pm}$ is approximately normal with mean β_0 and asymptotic variance consistently estimated by $\hat{\Gamma}_{pm}^{-1} \hat{\Omega}_{pm} \hat{\Gamma}_{pm}^{-1}$ where

$$\hat{\Gamma}_{pm}^{-1} = -\frac{\partial}{\partial \beta^T} \mathbb{P}_n \left\{ U_{pm}(L_{(R)}, R; \beta, \widetilde{\eta}) \right\} \Big|_{\widehat{\beta}_{pm}};$$

$$\hat{\Omega}_{pm} = n^{-1} \mathbb{P}_n \left[U_{pm}(L_{(R)}, R; \widehat{\beta}_{pm}, \widetilde{\eta}) + \frac{\partial}{\partial \eta^T} \mathbb{P}_n U_{pm}(L_{(R)}, R; \widehat{\beta}_{pm}, \eta) \Big|_{\widetilde{\eta}} \widehat{IF}_\eta \right]^{\otimes 2};$$

$$\widehat{IF}_\eta = -\left[\frac{\partial^2}{\partial \eta \partial \eta^T} \mathbb{P}_n \left\{ \text{llik}_{l,c}(\eta) \right\} \Big|_{\widetilde{\eta}} \right]^{-1} \frac{\partial}{\partial \eta} \left\{ \sum_{r \neq 1} \text{llik}_{l,c}(\eta) \right\} \Big|_{\widetilde{\eta}}.$$

4.3 Doubly robust and multiply robust LDCM estimation

We have now described two separate approaches for estimating the full data functional β_0 under the LDCM, IPW, and PM estimation, each of which depends on variation independent parameter of the joint distribution of $f(R, L)$ given in Lemma 2. Validity of IPW estimation relies on correct specification of the nonresponse model $\mathcal{M}_R$, while PM estimation relies for consistency on correct specification of $\mathcal{M}_L$. When L is sufficiently high dimensional, one cannot be confident that

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0325

either, if any, model is correctly specified, it is of interest to develop a doubly robust estimation approach that is guaranteed to deliver valid inferences about β_0 provided either $\mathcal{M}_R$ or $\mathcal{M}_L$ is correctly specified, but not necessarily both. We aim to develop a consistent estimator of β_0 in the semiparametric union model $\mathcal{M}_{DR} = \mathcal{M}_R \cup \mathcal{M}_L$.

To describe the DR approach, let

$$\begin{aligned} V(\beta,\alpha,\eta) &\equiv v\left(L_{(R)}, R; \beta, \alpha, \eta\right) \\ &= \left\{ \frac{1(R=1)}{\Pi_1(\alpha)} U(L;\beta) \right\} \\ &- \frac{1(R=1)}{\Pi_1(\alpha)} \sum_{r \neq 1} \Pi_r(\alpha) E\left[U(L;\beta) | L_{(r)}, R=1; \eta\right] \\ &+ \sum_{r \neq 1} I(R=r) E\left[U(L;\beta) | L_{(r)}, R=1; \eta\right] \end{aligned}$$

and let $\widehat{\beta}_{dr}$ be the solution to

$$0 = \mathbb{P}_n V\left(\widehat{\beta}_{dr}, \widetilde{\alpha}, \widetilde{\eta}\right). \tag{14}$$

Theorem 3 *If assumptions* (1) *and* (2) *hold with* F_ε *the extreme value distribution, then, under standard regularity conditions,* $\widehat{\beta}_{dr}$ *is consistent and asymptotically normal in the union model* $\mathcal{M}_{DR}$ *with asymptotic variance consistently estimated by* $\hat{\Gamma}_{dr}^{-1} \hat{\Omega}_{dr} \hat{\Gamma}_{dr}^{-1}$, *where*

$$\begin{aligned} \hat{\Gamma}_{dr}^{-1} &= -\frac{\partial}{\partial \beta^T} \mathbb{P}_n \left\{ V(\beta, \widetilde{\alpha}, \widetilde{\eta}) \right\} \Big|_{\widehat{\beta}_{dr}}; \\ \hat{\Omega}_{dr} &= n^{-1} \mathbb{P}_n \left[V\left(\widehat{\beta}_{dr}, \widetilde{\alpha}, \widetilde{\eta}\right) \right. \\ &\quad \left. + \frac{\partial}{\partial \eta^T} \mathbb{P}_n \left\{ V\left(\widehat{\beta}_{dr}, \widetilde{\alpha}, \eta\right) \right\} \Big|_{\widetilde{\eta}} \widehat{IF}_\eta + \frac{\partial}{\partial \alpha^T} \mathbb{P}_n \left\{ V\left(\widehat{\beta}_{dr}, \alpha, \widetilde{\eta}\right) \right\} \Big|_{\widetilde{\alpha}} \widehat{IF}_\alpha \right]^{\otimes 2}. \end{aligned}$$

This formally establishes the DR property of $\widehat{\beta}_{dr}$. Instead of these estimators of asymptotic variance, one can use the nonparametric bootstrap to obtain inferences based on either $\widehat{\beta}_{dr}$, $\widehat{\beta}_{ipw}$, or $\widehat{\beta}_{pm}$.

Remark 2 Equation (8) of Lemma 2 implies that $f(R=1|l)$ (which only depends on $\{\text{Odds}_r\left(l_{(r)}\right) : r\}$) and $f(l|R=1)$ are variation independent under the CCMV restriction. This variation inde-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0325

pendence is important as double robustness is meaningful only if it is possible *a priori* for both of the nuisance models to be correctly specified, see Robins and Rotnitzky (2001) and Richardson et al. (2016, Remark 3.1). However, in general, $f(l|r)$ and $f(r|l)$ are variation dependent even under CCMV.

It is possible to make the estimator $\widehat{\beta}_{dr}$ even more robust by a modification to estimation of the nuisance parameter η. Specifically, suppose that for each r, the conditional density $f\left(L_{(-r)}|L_{(r)}, r; \eta\right) = f\left(L_{(-r)}|L_{(r)}, r; \eta_r\right) = f\left(L_{(-r)}|L_{(r)}, R=1; \eta_r\right)$ only depends on the subset of parameter $\eta_r \subset \eta$, where there may be parameter overlap across patterns $\eta_r \cap \eta_{r'} \neq \varnothing$ for distinct patterns r and r'. Let $\mathcal{M}_L(r)$ be the semiparametric model that only specifies $f\left(L_{(-r)}|L_{(r)}, R=1; \eta_r\right)$, allowing the density of $f(L_{(r)}|R=1)$ and the missing data process to remain unspecified. Here $\mathcal{M}_L \subseteq \bigcap_{r \neq 1} \mathcal{M}_L(r)$. Let $\overline{\eta}_r$ denote the complete-case MLE under $\mathcal{M}_L(r)$: $\overline{\eta}_r = \arg\max_{\eta_r} \mathbb{P}_n I(R=1) f(L_{(-r)}|L_{(r)}, R=1; \eta_r)$. Likewise, let $\mathcal{M}_R(r)$ denote the semiparametric model that specifies the nonresponse model $\Pi_{r,c}(\alpha_r)$, and is otherwise unspecified. Then $\mathcal{M}_R = \bigcap_{r \neq 1} \mathcal{M}_R(r)$. Consider the pattern-specific union model $\mathcal{M}_{DR}(r) = \mathcal{M}_R(r) \cup \mathcal{M}_L(r)$, which is the set of laws with either $\mathcal{M}_R(r)$ or $\mathcal{M}_L(r)$ correctly specified. The intersection submodel of these laws $\mathcal{M}_{MR} = \bigcap_{r \neq 1} \mathcal{M}_{DR}(r) = \bigcap_{r \neq 1} \{\mathcal{M}_R(r) \cup \mathcal{M}_L(r)\}$ is the set of laws such that the union model for each r holds. Then $\mathcal{M}_{DR} \subseteq \mathcal{M}_{MR}$ since the first union model requires that either the entire nonresponse process is correctly specified, or the joint complete-case distribution of L is correctly specified; in contrast, $\mathcal{M}_{MR}$ requires only correct specification of one of the two models for each pattern. An estimator of β_0 that is consistent in model $\mathcal{M}_{MR}$ is said to be multiply-robust, or 2^J-robust (Vansteelandt et al (2007)) for a J non-monotone missing data patterns.

Corollary 4 *If assumptions* (1) *and* (2) *hold with F_ε the extreme value distribution, then, under standard regularity conditions, $\widehat{\beta}_{mr}$ is consistent and asymptotically normal in the union model $\mathcal{M}_{MR}$, where $\widehat{\beta}_{mr}$ is defined as $\widehat{\beta}_{dr}$ with $\overline{\eta}_r$ used to estimate η_r.*

This result describes an estimator with the MR property which states that given J nonresponse patterns, the analyst would in principle have (under our identifying assumptions) 2^J opportunities to obtain valid inferences about β_0. This is to be contrasted with the single chance to valid inferences

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0325

offered by IPW or PM approaches, or the two chances offered by the DR estimator. For inference, one can readily adapt the large sample variance estimator given in Theorem 3, or alternatively use the nonparametric bootstrap.

4.4 Simulation Study

We performed a simulation study to investigate the performance of the various estimators as described in finite samples. We generated 1000 samples of size $n = 2000$. We took independent and identically distributed (Y, X) generated from a normal mixture models: $(Y, X) \sim \sum_{k=1}^{3} \pi_k N(\mu_k, \Sigma)$, where $\pi_1 = 1/2, \pi_2 = e/(2+2e), \pi_3 = 1/(2+2e), \mu_1 = (0,0)^T, \mu_2 = (1,1)^T, \mu_3 = (1,2)^T$ and $\Sigma = (\sigma_{ij})$, where $\sigma_{11} = \sigma 12 = 1, \sigma_{22} = 2$. We considered four missing data patterns $L_{(R)}$: $L_{(1)} = L$, $L_{(2)} = X$, $L_{(3)} = Y$, $L_{(4)} = \varnothing$. Conditional on the generated full data, the missing data pattern was then generated under the mechanism

$$
\begin{aligned}
P(R = 1 \mid X, Y) &= \frac{1}{1 + \exp(X) + \exp(2Y) + \exp(-1)}; \\
P(R = 2 \mid X, Y) &= \frac{\exp(X)}{1 + \exp(X) + \exp(2Y) + \exp(-1)}; \\
P(R = 3 \mid X, Y) &= \frac{\exp(2Y)}{1 + \exp(X) + \exp(2Y) + \exp(-1)}; \\
P(R = 4 \mid X, Y) &= \frac{\exp(-1)}{1 + \exp(X) + \exp(2Y) + \exp(-1)}.
\end{aligned}
$$

Since, for each missing data pattern r, $P(R = r \mid X, Y)$ depends on the full data (X, Y), the missing data mechanism is MNAR. The identifiability of normal mixture models in the MNAR setting has previously been considered in Miao et al.. (2016). The full data target parameter of interest is $\beta = E(Y) = \sum_r p_r E[Y|R = r] = (2 + \exp(1))/(2 + 2\exp(1))$, with full data estimating equation $U(\beta) = Y - \beta$.

We implemented Little's PM approach as well as our IPW and DR estimators. In doing so, correct specification of the nonresponse process entailed matching the data-generating mechanism described above, $\text{Odds}_2\left(L_{(2)}\right) = \alpha_{20} + \alpha_{21} X$, $\text{Odds}_3\left(L_{(3)}\right) = \alpha_{30} + \alpha_{31} Y$, $\text{Odds}_4\left(L_{(4)}\right) = \alpha_{40}$. Misspecification of these models occured by instead fitting $\text{Odds}_2\left(L_{(2)}\right) = \alpha_{20} + \alpha_{21} X^2$ and

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0325

Table 1: Monte Carlo results of the IPW, PM and DR estimators: accuracy of standard deviation estimator and coverage probabilities. The sample size is 2000

	bth*	nrm	ccm	bad
Estimated SD / Monte Carlo SD				
IPW	0.951	0.951	0.438	0.438
PM	0.993	0.979	0.993	0.979
DR	0.995	0.995	0.886	0.725
Estimated SD / Bootstrapped SD				
IPW	0.994	0.994	0.932	0.932
PM	1.000	1.002	1.000	1.002
DR	0.999	0.990	0.973	0.951
Coverage**				
IPW	0.938	0.938	0.080	0.080
PM	0.954	0.001	0.954	0.001
DR	0.948	0.947	0.953	0.030

*: `bth`: both models correct; `nrm`: nonresponse model correct; `ccm`: complete-case model correct; `bad`: both models incorrect.
**: Nominal level = 95%.

$\text{Odds}_3\left(L_{(3)}\right) = \alpha_{30} + \alpha_{31}Y^2$. Likewise, correct specification for the PM approach entailed defining $E\left(Y|R=2,X\right) = E\left(Y|R=1,X\right) = \gamma_{20} + \gamma_{21}X$, while the incorrect model $E\left(Y|R=1,X\right) = \gamma_{20} + \gamma_{21}X^2$ was used to assess the impact of model mis-specification of the complete-case distribution. As $U\left(\beta\right)$ does not depend on $X, E\left[U\left(\beta\right)|R=3,L_{(3)}\right] = U\left(\beta\right)$. We explored four scenarios corresponding to (1) correct $f\left(R|L\right)$ and $f(L|R=1)$, (2) correct $f\left(R|L\right)$ but incorrect $f(L|R=1)$; (3) correct $f(L|R=1)$ but incorrect $f\left(R|L\right)$; (4) incorrect $f\left(R|L\right)$ and $f(L|R=1)$.

Results in Table 1 confirm our theoretical results, and show that, as expected, IPW has small bias in scenarios (1) and (2) only, PM has small bias in scenarios (1) and (3), and DR has small bias in scenarios (1)-(3). In scenario (4) where all models are incorrect, as expected all estimators are significantly biased. When, as in the first scenario, model misspecification is absent, IPW has larger root mean squared error (RMSE) than PM, but DR is comparable to PM, at least in this simulation setting. The RMSE of DR follows closely that of PM in scenarios (1) and (3), suggesting that the potential efficiency loss incurred to obtain DR inference relative to PM inference may not be substantial in practice. Table 1 of the Supplemental Appendix summarizes simulation

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0325

results assessing the performance of our estimators of asymptotic variance and coverage of Wald confidence intervals using estimated standard errors for the three estimators under consideration. The results largely indicate that our standard error estimators are consistent in all scenarios where the point estimators are also consistent, including under partial model misspecification for the DR estimator (see the comparison to Monte Carlo standard errors in Table 1 of the Supplemental Appendix). However, our standard error estimators appear to break down severely whenever model mis-specification induces bias in parameter estimates. The performance of the nonparametric bootstrap closely follows that of our estimators in all instances, and also appears to break down under bias inducing model misspecification. We do not view this as a serious limitation given that inferences are in such cases unreliable, even with a consistent estimator of standard error.

4.5 A data application

The empirical application concerns a study of the association between maternal exposure to highly active antiretroviral therapy (HAART) during pregnancy and birth outcomes among HIV-infected women in Botswana. A detailed description of the study cohort is in Chen et al. (2012). The entire study cohort consists of 33148 obstetrical records abstracted from 6 sites in Botswana for 24 months. Our current analysis focuses on the subset of women who were known to be HIV positive (n = 9711). The birth outcome of interest is preterm delivery, defined as delivery < 37 weeks gestation. 6.7% of the outcomes were not observed. The data also contain the risk factors of interest that are also subject to missingness (Table 2): whether CD4+ cell count is less than 200 cells/μL, and whether a woman continued HAART from before pregnancy or not.

Our goal is to correlate these factors with preterm delivery using a logistic regression: the parameter of interest is the vector of coefficients of the corresponding logistic regression. We implemented the complete-case (CC) analysis, in addition to the LDCM IPW, PM and DR estimators. Estimation of the nonresponse process used the fairly generic specification $\log \text{Odds}_r \left(L_{(r)}; \alpha_r\right) = \alpha_r' q_r \left(L_{(r)}\right)$, where $q_r \left(L_{(r)}\right)$ included all main effects and two-way interactions of components of $L_{(r)}$ while PM specified the log-linear model $\Pr(L|R=1) \propto \exp\{\eta' L\}$.

Table 3 summarizes results for the complete analysis (CC), together with Little's PM analysis

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0325

Table 2: Data analysis: tabulation of missing data patterns. The total sample size is 9711. Missing variables are coded by 0. The first row represents the complete case

Pattern (R)	Preterm Delivery	Low CD4 Count	Cont. HAART	percentage
1	1	1	1	10.5%
2	0	1	1	0.7%
3	1	0	1	18.3%
4	0	0	1	1.6%
5	1	1	0	33.9%
6	0	1	0	1.5%
7	1	0	0	30.6%
8	0	0	0	2.9%

Table 3: Data analysis: estimated odds ratios of preterm delivery associated with various risk factors. The 95% confidence intervals are estimated based on bootstrap samples

	Low CD4 Count	Cont HAART
CC	0.782 (0.531,1.135)	1.142 (0.810,1.620)
IPW	0.924 (0.631,1.338)	1.180 (0.847,1.638)
PM	0.963 (0.704,1.318)	1.175 (0.881,1.598)
DR	1.020 (0.742,1.397)	1.158 (0.869,1.560)

and our two semiparametric estimators (IPW and DR). The results suggest that the association between CD4 count and preterm delivery may be subject to selection bias to a greater extent than that of HAART and preterm delivery. The estimated odds ratio for CD4 count is about 20% larger for IPW, PM, and DR compared to the CC odds ratio, whereas the odds ratio for HAART is quite similar for all four estimators. Although PM generally appears less variable, there are no notable differences between inferences obtained using IPW, PM or DR, providing no evidence that either IPW or PM, might be subject to misspecification bias.

5 Inference for general DCM

Consider a DCM with user-specified F_ε, a well-defined continuous CDF. Local identification under assumption (3) is best understood with discrete data. In this vein, suppose that $L_{(r)}$ takes on M_r levels, so $\Delta\mu_{1r}\left(L_{(r)}\right)$ depends on at most M_r unknown parameters, but for user-supplied M_r-dimensional function $G_r = g_r\left(L_{(r)}\right)$. Let $W_r\left(G_r\right) = G_r \times \left[1\left\{R = r\right\} - 1\left\{R = 1\right\}\Pi_r/\Pi_1\right]$. It is

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0325

straightforward to verify that

$$E\left\{W_r\left(G_r\right)\right\} = 0 \text{ for } r = 2, ... \tag{15}$$

yielding the M_r restrictions needed to identify each $\Delta\mu_{1r}$. Naturally, components of G_r should be chosen appropriately to avoid redundancy and linear dependence. A similar argument could in principle be crafted to establish local identification if L contains continuous components. This is not further pursued in this paper. Equation (15) motivates a simple approach for estimating Π_r in practice. Suppose that one posits a parametric model $\Delta\mu_{1r}\left(L_{(r)};\alpha_r\right)$ for $\Delta\mu_{1r}\left(L_{(r)}\right)$ with finite dimensional unknown parameter α_r, for all r. Then, the empirical version of (15) would in principle deliver an estimator $\widehat{\alpha}=\{\widehat{\alpha}_r : r\}$ of $\alpha = \{\alpha_r : r\}$,

$$\mathbb{P}_n\left\{W_r\left(\widehat{G}_r;\widehat{\alpha}\right)\right\} = 0 \text{ for } r = 2, ...$$

where $W_r\left(\widehat{G}_r;\widehat{\alpha}\right) = \widehat{G}_r \times \left[1\left\{R = r\right\} - 1\left\{R = 1\right\}\Pi_r\left(\widehat{\alpha}\right)/\Pi_1\left(\widehat{\alpha}\right)\right]$. A convenient choice for $\widehat{G}_r = \partial\Delta\mu_{1r}\left(L_{(r)};\widehat{\alpha}_r\right)/\partial\widehat{\alpha}_r$. Under mild regularity conditions, $\widehat{\alpha}$ is consistent and asymptotically normal provided $\Delta\mu_{1r}\left(L_{(r)};\alpha_r\right)$ is correctly specified for all r.

Given a consistent estimator of Π_1, IPW inferences about β_0 can be obtained as previously described. Likewise, maximum likelihood estimation is straightforward by maximizing the model for the likelihood given in Lemma 1. Unfortunately, outside of the LDCM, to the best of our knowledge, it does not appear possible to obtain DR and MR inferences for DCMs.

This analysis requires evaluation of the integral defining Π_r.Thus, let

$$Q_r\left(\varepsilon\right) = \prod_{s\neq r} F_\varepsilon\left(\Delta\mu_{1s}\left(L_{(s)}\right) - \Delta\mu_{1r}\left(L_{(r)}\right) + \varepsilon\right).$$

A reliable approximation of $\Pi_r = \int Q_r\left(\varepsilon\right) f_\varepsilon\left(\varepsilon\right) d\varepsilon$ can be achieved numerically by Gauss-Hermite Quadrature (Liu and Pierce (1994)). For instance, if f_ε is standard normal, then the approximate Gaussian Discrete Choice Model is given by $\Pi_r \approx \sum_{m=1}^{M} Q_r\left(\varepsilon_m\right) w_m$, where the nodes ε_m are the zeroes of the mth order Hermite polynomial and the w_m are suitably defined weights (Davis and Rabinowitz (1975)).

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0325

6 Conclusion

In this paper, we have described the DCM as an all-purpose, flexible, and easy-to-implement general class of models for nonmonotone nonignorable nonresponse. The LDCM has several advantages including giving rise to four distinct strategies for inference: IPW, PM, DR, and MR estimation. Simulation studies and an application suggest good finite sample performance of IPW, PM, and DR estimation; although not directly evaluated, we expect the same to apply to MR estimation.

Identification conditions such as CCMV are not empirically testable and therefore, it is important that inferences are assessed for sensitivity to violation of such assumptions. Such an approach for sensitivity analysis for violation of CCMV restriction is outlined in the Supplemental Appendix.

7 Supplementary Materials

The supplementary materials include an outline of sensitivity analysis for CCMV, proof of Lemmas as well as additional simulation results.

References

[1] Albert, P. S. (2000). A transitional model for longitudinal binary data subject to nonignorable missing data. Biometrics **56**, 602-608.

[2] Andridge, R. R. and Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. International Statistical Review **78**, 40-64.

[3] Chen, H. Y. (2007). A semiparametric odds ratio model for measuring association. Biometrics **63**, 413-421.

[4] Chen, J. Y., Ribaudo, H. J., Souda, S., Parekh, N., Ogwu, A., Lockman, S., Powis, K., Dryden-Peterson, S., Creek, T., Jimbo, W., Madidimalo, T., Makhema, J., Essex,M. and Shapiro, R. L. (2012). Highly active antiretroviral therapy and adverse birth outcomes among hiv-infected women in botswana. The Journal of Infectious Diseases **206**, 1695-1705.

[5] Davis, P. J. & Rabinowitz, P. (1975). Methods of Numerical Integration. New York: Academic Press.

[6] Deltour, I., Richardson, S. and Le Hesran J. Y. (1999). Stochastic algorithms for Markov models estimation with intermittent missing data. Biometrics **55**, 565-573.

[7] 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) **39**, 1-38.

[8] Fairclough, D. L., Peterson, H. F., Cella, D., & Bonomi, P. (1998). Comparison of several model-based methods for analysing incomplete quality of life data in cancer clinical trials. Statistics in Medicine **17**, 781-796.

[9] Horton, N. J. and Laird, N. M. (1999). Maximum likelihood analysis of generalized linear models with missing covariates. Statistical Methods in Medical Research **8**, 37-50.

[10] Horton, N. J. and Lipsitz, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. The American Statistician **55**, 244-254.

[11] Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association **47**, 663-685.

[12] Ibrahim, J. G. and Chen, M. H. (2000). Power prior distributions for regression models. Statistical Science **15**, 46-60.

[13] Ibrahim, J. G., Chen, M. H. and Lipsitz, S. R. (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. Biometrika **88**, 551-564.

[14] Ibrahim, J. G., Chen, M. H. and Lipsitz, S. R. (2002). Bayesian methods for generalized linear models with covariates missing at random. Canadian Journal of Statistics **30**, 55-78.

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

[16] Little, R.J., 1993. Pattern-mixture models for multivariate incomplete data. Journal of the American Statistical Association **88**, 125-134.

[17] Little, R. J. and Rubin, D. B. (2002). Statistical Analysis with Missing Data. Wiley.

[18] Liu, Q., & Pierce, D. A. (1994). A note on Gauss—Hermite quadrature. Biometrika **81**, 624-629.

[19] McFadden, D. L. (1984). Econometric analysis of qualitative response models. Handbook of Econometrics, Volume II. Chapter 24. Elsevier Science Publishers BV.

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

[21] Richardson, T.S., Robins, J.M. and Wang, L., 2016. On Modeling and Estimation for the Relative Risk and Risk Difference. Journal of the American Statistical Association, (just-accepted).

[22] 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**, 846-866.

[23] Robins, J. M. and Gill, R. D. (1997). Non-response models for the analysis of non-monotone ignorable missing data. Statistics in Medicine **16**, 39-56.

[24] Robins JM. (1997). Non-response models for the analysis of non-monotone non-ignorable missing data. Statistics in Medicine **16**, 21-37.

[25] Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. Statistics in Medicine **16**, 285-319.

[26] Robins JM, Rotnitzky A, Scharfstein D. (1999). 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. Halloran, M.E. and Berry, D., eds. IMA Volume 116, NY: Springer-Verlag, pp. 1-92.

[27] Robins JM, Rotnitzky A. (2001). Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer". Statistica Sinica **11**, 920-936.

[28] Rubin, D. B. (1976). Inference and missing data. Biometrika **63**, 581-592.

[29] Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. Journal of the American Statistical Association **72**, 538-543.

[30] Schafer, J. (1997). Analysis of Incomplete Multivariate Data. Chapman and Hall.

[31] Sun, BL. and Tchetgen Tchetgen, E. J. (2016), On Inverse Probability Weighting for Non-monotone Missing at Random Data, Journal of the American Statistical Association. Advance online publication. doi:10.1080/01621459.2016.1256814

[32] Tchetgen Tchetgen, E. J., Robins, J. M., & Rotnitzky, A. (2010). On doubly robust estimation in a semiparametric odds ratio model. Biometrika **97**, 171-180.

[33] Train, K. (2009). Discrete Choice Methods with Simulation. Cambridge University Press.

[34] Troxel, A. B., Harrington, D. P., & Lipsitz, S. R. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. Journal of the Royal Statistical Society: Series C (Applied Statistics) **47**, 425-438.

[35] Troxel, A. B., Lipsitz, S. R. and Harrington, D. P. (1998), Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data. Biometrika **85**, 661-672.

[36] Tsiatis, A. (2006). Semiparametric Theory and Missing Data. Springer.

[37] van der Vaart, A. (1998). Asymptotic Statistics. Cambridge University Press.

[38] Vansteelandt S, Rotnitsky A, Robins JM. (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. Biometrika **94**, 841-860.

[39] Zhou, Y., Little, R. J., & Kalbfleisch, J. D. (2010). Block-conditional missing at random models for missing data. Statistical Science **25**, 517-532.