Statistica Sinica **28** (2018), 1741-1759
doi:https://doi.org/10.5705/ss.202016.0328

SEQUENTIAL IDENTIFICATION OF NONIGNORABLE MISSING DATA MECHANISMS

Mauricio Sadinle and Jerome P. Reiter

[1] *University of Washington and* [2] *Duke University*

Abstract: With nonignorable missing data, likelihood-based inference should be based on the joint distribution of the study variables and their missingness indicators. These joint models cannot be estimated from the data alone, thus requiring the analyst to impose restrictions that make the models uniquely obtainable from the distribution of the observed data. We present an approach for constructing classes of identifiable nonignorable missing data models. The main idea is to use a sequence of carefully set up identifying assumptions, whereby we specify potentially different missingness mechanisms for different blocks of variables. We show that the procedure results in models with the desirable property of being non-parametric saturated.

Key words and phrases: Identification, missing not at random, non-parametric saturation, partial ignorability, sensitivity analysis.

1. Introduction

When data are missing not at random (MNAR) (Rubin (1976)), appropriate likelihood-based inference requires explicit models for the *full-data distribution*, i.e., the joint distribution of the study variables and their missingness indicators. Because of the missing data, this distribution is not uniquely identified from the observed data alone (Little and Rubin (2002)). To enable inference, analysts must impose restrictions on the full-data distribution. Such assumptions generally are untestable; however, a minimum desideratum is that they result in a unique full-data distribution for the *observed-data distribution* at hand, i.e., the distribution that can be identified from the incomplete data.

We present a strategy for constructing identifiable full-data distributions with nonignorable missing data. In its most general form, the strategy is to expand the observed-data distribution sequentially by identifying parts of the full-data distribution associated with blocks of variables, one block at a time. This partitioning of the variables allows analysts to specify different missingness mechanisms in the different blocks; for example, use the missing at random

(MAR, Rubin (1976)) assumption for some variables and a nonignorable missingness assumption for the rest to obtain a partially ignorable mechanism (Harel and Schafer (2009)). We ensure that the resulting full-data distributions are *non-parametric saturated* (NPS, Robins (1997)), that is, their implied observed-data distribution matches the actual observed-data distribution, as detailed in Section 2.2.

Related approaches to partitioning variables with missing data have appeared previously in the literature. Zhou, Little and Kalbfleisch (2010) proposed to model blocks of study variables and their missingness indicators in a sequential manner; however, their approach does not guarantee identifiability of the full-data distribution. Harel and Schafer (2009) mentioned the possibility of treating the missingness in blocks of variables differently, but provide no results on identification. Robins (1997) proposed the group permutation missingness mechanism, which assumes MAR sequentially for blocks of variables and results in a NPS model. This is a particular case of our more general procedure, as we describe in Section 3.4.

The remainder of the article is organized as follows. In Section 2, we describe notation and provide more details on the NPS property. In Section 3, we introduce our strategy for identifying a full-data distribution in a sequential manner. In Section 4 we present some examples of how to use this strategy for the case of two categorical study variables, for the construction of partially ignorable mechanisms, and for sensitivity analyses. Finally, in Section 5 we discuss possible future uses of our identifying approach.

2. Notation and Background

2.1. Notation

Let $\mathbf{X} = (X_1, \ldots, X_p)$ denote p random variables taking values on a sample space $\mathcal{X}$. Let M_j be the missingness indicator for variable j, where $M_j = 1$ when X_j is missing and $M_j = 0$ when X_j is observed. Then $\mathbf{M} = (M_1, \ldots, M_p)$ takes values on $\{0,1\}^p$. Let μ be a dominating measure for the distribution of $\mathbf{X}$, and let ν represent the product measure between μ and the counting measure on $\{0,1\}^p$. The full-data distribution is the joint distribution of $\mathbf{X}$ and $\mathbf{M}$. We call its density f with respect to ν the *full-data density*. The full-data distribution cannot be recovered from sampled data, even with an infinite sample size.

An element $\mathbf{m} = (m_1, \ldots, m_p) \in \{0,1\}^p$ is called a *missingness pattern*. Given $\mathbf{m} \in \{0,1\}^p$ we define $\bar{\mathbf{m}} = \mathbf{1}_p - \mathbf{m}$ to be the indicator vector of observed

variables, where $\mathbf{1}_p$ is a vector of ones of length p. For each $\mathbf{m}$, we define $\mathbf{X}_\mathbf{m} = (X_j : m_j = 1)$ to be the missing variables and $\mathbf{X}_{\bar{\mathbf{m}}} = (X_j : \bar{m}_j = 1)$ to be the observed variables, which have sample spaces $\mathcal{X}_\mathbf{m}$ and $\mathcal{X}_{\bar{\mathbf{m}}}$, respectively. The observed-data distribution is the distribution involving the observed variables and the missingness indicators, which has density $f(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{M} = \mathbf{m}) = \int_{\mathcal{X}_\mathbf{m}} f(\mathbf{X} = \mathbf{x}, \mathbf{M} = \mathbf{m})\mu(d\mathbf{x}_\mathbf{m})$, where $\mathbf{x} \in \mathcal{X}$ represents a generic element of the sample space, and we define $\mathbf{x}_\mathbf{m}$ and $\mathbf{x}_{\bar{\mathbf{m}}}$ similarly as with the random vectors.

An alternative way of representing the observed-data distribution is by introducing the *materialized variables* $\mathbf{X}^* = (X_1^*, \ldots, X_p^*)$, where

$$X_j^* \equiv \begin{cases} X_j, & \text{if } M_j = 0, \\ *, & \text{if } M_j = 1, \end{cases}$$

and "$*$" is a placeholder for missingness. The sample space $\mathcal{X}_j^*$ of each X_j^* is the union of $\{*\}$ and the sample space $\mathcal{X}_j$ of X_j. The materialized variables contain all the observed information: if $X_j^* = *$ then X_j was not observed, and if $X_j^* = x_j$ for any value $x_j \neq *$ then X_j was observed and $X_j = x_j$. Given $\mathbf{m} \in \{0,1\}^p$ and $\mathbf{x}_{\bar{\mathbf{m}}} \in \mathcal{X}_{\bar{\mathbf{m}}}$, we define $\mathbf{x}^* \equiv \mathbf{x}^*(\mathbf{m}, \mathbf{x}_{\bar{\mathbf{m}}})$, such that $\mathbf{x}^*_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}$ and $\mathbf{x}^*_\mathbf{m} = \boldsymbol{*}$, where $\boldsymbol{*}$ is a vector with the appropriate number of $*$ symbols. For example, if $\mathbf{m} = (1,1,0)$ and $\mathbf{x}_{\bar{\mathbf{m}}} = x_3$, then $\mathbf{x}^* = (*, *, x_3)$. The event $\mathbf{X}^* = \mathbf{x}^*(\mathbf{m}, \mathbf{x}_{\bar{\mathbf{m}}})$ is equivalent to $\mathbf{M} = \mathbf{m}$ and $\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}$, which implies that the distribution of $\mathbf{X}^*$ is equivalent to the observed-data distribution. Therefore, with some abuse of notation, the observed-data density can be written in terms of $\mathbf{X}^*$, that is $f\{\mathbf{X}^* = \mathbf{x}^*(\mathbf{m}, \mathbf{x}_{\bar{\mathbf{m}}})\} \equiv f(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{M} = \mathbf{m})$. When there is no need to refer to the $\mathbf{m}$ and $\mathbf{x}_{\bar{\mathbf{m}}}$ that define $\mathbf{x}^*$, we simply write $f(\mathbf{X}^* = \mathbf{x}^*)$ to denote the observed-data density evaluated at an arbitrary point.

In what follows we often write $f(\mathbf{X} = \mathbf{x}, \mathbf{M} = \mathbf{m})$ simply as $f(\mathbf{X}, \mathbf{M})$, $f(\mathbf{X}^* = \mathbf{x}^*)$ as $f(\mathbf{X}^*)$, and likewise for other expressions, provided that there is no ambiguity. For the sake of simplicity, we use "f" for technically different functions, but their actual interpretations should be clear from the arguments passed to them. For example, we denote the *missingness mechanism* as $f(\mathbf{M} = \mathbf{m}|\mathbf{X} = \mathbf{x})$, or simply $f(\mathbf{M}|\mathbf{X})$.

2.2. Non-parametric saturated modeling

Since the true joint distribution of $\mathbf{X}$ and $\mathbf{M}$ cannot be identified from observed data alone, we need to work under the assumption that the full-data distribution falls within a class defined by a set of restrictions.

Definition 1 (Identifiability)**.** *Consider a class of full-data distributions* $\mathcal{F}_A$

defined by a set of restrictions A. We say that the class $\mathcal{F}_A$ is identifiable if there is a mapping from the set of observed-data distributions to $\mathcal{F}_A$.

If we only require identifiability from a set of full-data distributions, two different observed-data distributions could map to the same full-data distribution. Robins (1997) introduced the stricter concept of a class of full-data distributions being non-parametric saturated – also called non-parametric identified (Vansteelandt et al. (2006); Daniels and Hogan (2008)).

Definition 2 (Non-parametric Saturation)**.** *Consider a class of full-data distributions $\mathcal{F}_A$ defined by a set of restrictions A. We say that the class $\mathcal{F}_A$ is non-parametric saturated if there is a* one-to-one *mapping from the set of observed-data distributions to $\mathcal{F}_A$.*

The set A of restrictions, or identifiability assumptions, that define a NPS class allow us to build a full-data distribution, say with density $f_A(\mathbf{X} = \mathbf{x}, \mathbf{M} = \mathbf{m})$, from an observed-data distribution with density $f(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{M} = \mathbf{m})$, so that $f_A(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{M} = \mathbf{m}) = f(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{M} = \mathbf{m})$, where by definition $f_A(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{M} = \mathbf{m}) = \int_{\mathcal{X}_{\mathbf{m}}} f_A(\mathbf{X} = \mathbf{x}, \mathbf{M} = \mathbf{m})\mu(d\mathbf{x}_{\mathbf{m}})$. In terms of $\mathbf{X}^*$, the NPS property is expressed as $f_A(\mathbf{X}^*) = f(\mathbf{X}^*)$.

NPS is a desirable property, particularly for comparing inferences under different approaches to handling nonignorable missing data. When two missing data models satisfy NPS, we can be sure that any discrepancies in inferences are due entirely to the different assumptions on the non-identifiable parts of the full-data distribution. In contrast, without NPS, it can be difficult to disentangle what parts of the discrepancies are due to the identifying assumptions and what parts are due to differing constraints on the observed-data distribution. Thus, NPS greatly facilitates sensitivity analysis (Robins (1997)).

For a given $\mathbf{m}$, we refer to the conditional distribution of the missing study variables given the observed data as the *missing-data distribution* – also known as the extrapolation distribution (Daniels and Hogan (2008)) – with density $f(\mathbf{X}_{\mathbf{m}} = \mathbf{x}_{\mathbf{m}} | \mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{M} = \mathbf{m})$. These distributions correspond to the non-identifiable parts of the full-data distribution. A NPS approach is equivalent to a recipe for building these distributions from the observed-data distribution without imposing constraints on the latter.

NPS models can be constructed in many ways. For example, in pattern mixture models, one can use the complete-case missing-variable restriction (Little (1993)), which sets $f(\mathbf{X}_{\mathbf{m}} = \mathbf{x}_{\mathbf{m}} | \mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{M} = \mathbf{m}) = f(\mathbf{X}_{\mathbf{m}} = \mathbf{x}_{\mathbf{m}} | \mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{M} = \mathbf{0}_p)$, for all $\mathbf{m} \in \{0, 1\}^p$. Although Little (1993) considered paramet-

ric models for each $f(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}|\mathbf{M} = \mathbf{m})$, this does not have to be the case, and therefore pattern mixture models can be NPS. Another example is the permutation missingness model of Robins (1997), which for a specific ordering of the study variables assumes that the probability of observing the kth variable depends on the previous study variables and the subsequent observed variables. The group permutation missingness model of Robins (1997) is an analog of the latter for groups of variables and is also NPS. Sadinle and Reiter (2017) introduced a missingness mechanism where each variable and its missingness indicator are conditionally independent given the remaining variables and missingness indicators, which leads to a NPS model. Tchetgen Tchetgen, Wang and Sun (2016) proposed a NPS approach based on discrete choice models. Finally, we note that MAR models also can be NPS, as shown by Gill, van der Laan and Robins (1997).

3. Sequential Identification Strategy

We consider the p variables as divided into K blocks, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_K)$, where $\mathbf{X}_k = (X_{t_{k-1}+1}, \ldots, X_{t_k})$. As our results only concern the identification of full-data distributions starting from an observed-data distribution, we assume that $f(\mathbf{X}^*) = f(\mathbf{X}_1^*, \ldots, \mathbf{X}_K^*)$ is known. The identification strategy consists of specifying a sequence of assumptions $A_1, \ldots, A_K$, one for each block of variables, where each A_k allows us to identify the conditional distribution of $\mathbf{X}_k$ and $\mathbf{M}_k$ given $\mathbf{X}_{<k} \equiv (\mathbf{X}_1, \ldots, \mathbf{X}_{k-1})$, $\mathbf{X}_{>k}^* \equiv (\mathbf{X}_{k+1}^*, \ldots, \mathbf{X}_K^*)$, and a carefully chosen subset of the missingness indicators $\mathbf{M}_{<k} \equiv (\mathbf{M}_1, \ldots, \mathbf{M}_{k-1})$ described below. We first provide a general description of how $A_1, \ldots, A_K$ allow us to identify parts of the full-data distribution in a sequential manner and then, in Theorem 1, present the formal identification result.

3.1. Description

We now present the steps needed to implement the identification strategy. A graphical summary of the procedure is provided in Figure 1.

Step 1. Write $f(\mathbf{X}^*) = f(\mathbf{X}_1^*|\mathbf{X}_{>1}^*)f(\mathbf{X}_{>1}^*)$. Consider an identifiability assumption A_1 on the distribution of $\mathbf{X}_1$ and $\mathbf{M}_1$ given $\mathbf{X}_{>1}^*$ to obtain a distribution with density $f_{A_1}(\mathbf{X}_1, \mathbf{M}_1|\mathbf{X}_{>1}^*)$ with the NPS property $f_{A_1}(\mathbf{X}_1^*|\mathbf{X}_{>1}^*) = f(\mathbf{X}_1^*|\mathbf{X}_{>1}^*)$. From this define $f_{A_1}(\mathbf{X}_1, \mathbf{M}_1, \mathbf{X}_{>1}^*) \equiv f_{A_1}(\mathbf{X}_1, \mathbf{M}_1|\mathbf{X}_{>1}^*)f(\mathbf{X}_{>1}^*)$.

Step 2. Divide the t_1 variables in $\mathbf{X}_1$ into two sets indexed by R_1 and S_1, where $R_1 \cup S_1 = \{1, \ldots, t_1\}$ and $R_1 \cap S_1 = \emptyset$. Let $\mathbf{M}_{R_1}$ and $\mathbf{M}_{S_1}$ be the corresponding missingness indicators, and write $f_{A_1}(\mathbf{X}_1, \mathbf{M}_1, \mathbf{X}_{>1}^*) = f_{A_1}(\mathbf{M}_{S_1}|\mathbf{X}_1, \mathbf{M}_{R_1}, \mathbf{X}_{>1}^*)f_{A_1}(\mathbf{X}_1, \mathbf{M}_{R_1}, \mathbf{X}_{>1}^*)$, where $f_{A_1}(\mathbf{X}_1, \mathbf{M}_{R_1}, \mathbf{X}_{>1}^*) = f_{A_1}(\mathbf{X}_2^*|\mathbf{X}_1, \mathbf{M}_{R_1},$

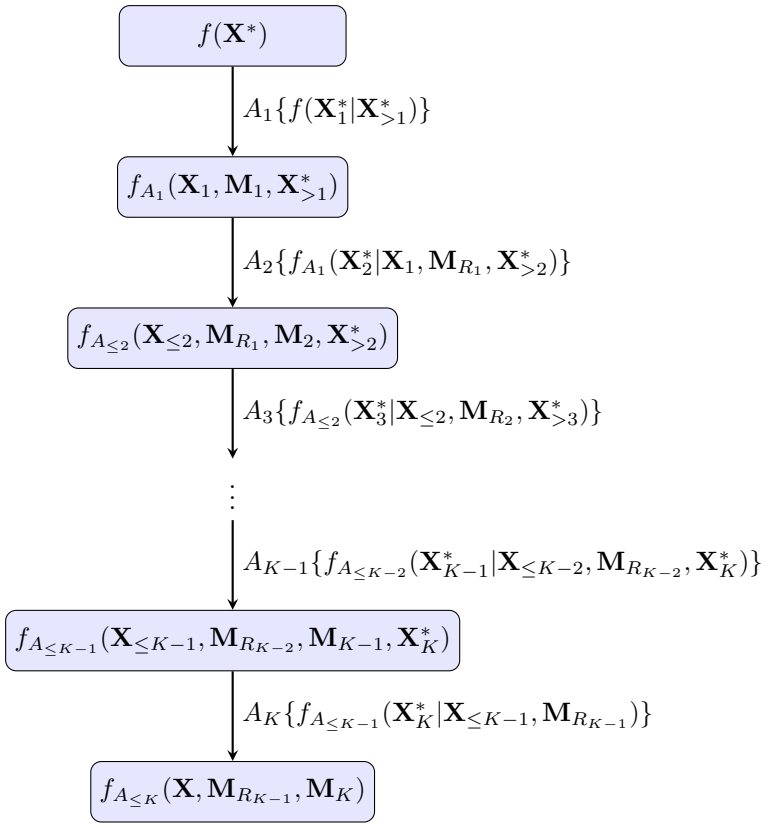

Figure 1. Sequential identification strategy. We write $A_k\{f_{A_{\leq k-1}}(\mathbf{X}_k^*|\cdots)\}$ to indicate that assumption A_k is being used to obtain a conditional full-data density $f_{A_{\leq k}}(\mathbf{X}_k, \mathbf{M}_k|\cdots)$ from $f_{A_{\leq k-1}}(\mathbf{X}_k^*|\cdots)$.

$\mathbf{X}_{>2}^*) f_{A_1}(\mathbf{X}_1, \mathbf{M}_{R_1}, \mathbf{X}_{>2}^*)$. Consider an identifiability assumption A_2 on the distribution of $\mathbf{X}_2$ and $\mathbf{M}_2$ given $\mathbf{X}_1, \mathbf{M}_{R_1}$ and $\mathbf{X}_{>2}^*$ to obtain a distribution with density $f_{A_{\leq 2}}(\mathbf{X}_2, \mathbf{M}_2|\mathbf{X}_1, \mathbf{M}_{R_1}, \mathbf{X}_{>2}^*)$ with the NPS property $f_{A_{\leq 2}}(\mathbf{X}_2^*|\mathbf{X}_1, \mathbf{M}_{R_1}, \mathbf{X}_{>2}^*)$ $= f_{A_1}(\mathbf{X}_2^*|\mathbf{X}_1, \mathbf{M}_{R_1}, \mathbf{X}_{>2}^*)$. Define

$$f_{A_{\leq 2}}(\mathbf{X}_{\leq 2}, \mathbf{M}_{R_1}, \mathbf{M}_2, \mathbf{X}_{>2}^*) \equiv f_{A_{\leq 2}}(\mathbf{X}_2, \mathbf{M}_2|\mathbf{X}_1, \mathbf{M}_{R_1}, \mathbf{X}_{>2}^*) f_{A_1}(\mathbf{X}_1, \mathbf{M}_{R_1}, \mathbf{X}_{>2}^*).$$

Step $k+1$. Take $f_{A_{\leq k}}(\mathbf{X}_{\leq k}, \mathbf{M}_{R_{k-1}}, \mathbf{M}_k, \mathbf{X}_{>k}^*)$ from the kth step. Let $R_k \cup S_k = \{t_{k-1}+1, \ldots, t_k\} \cup R_{k-1}$, $R_k \cap S_k = \emptyset$, and $\mathbf{M}_{R_k}$ and $\mathbf{M}_{S_k}$ be the corresponding missingness indicators. Write $f_{A_{\leq k}}(\mathbf{X}_{\leq k}, \mathbf{M}_{R_{k-1}}, \mathbf{M}_k, \mathbf{X}_{>k}^*) =$ $f_{A_{\leq k}}(\mathbf{M}_{S_k}|\mathbf{X}_{\leq k}, \mathbf{M}_{R_k}, \mathbf{X}_{>k}^*) f_{A_{\leq k}}(\mathbf{X}_{\leq k}, \mathbf{M}_{R_k}, \mathbf{X}_{>k}^*)$, where $f_{A_{\leq k}}(\mathbf{X}_{\leq k}, \mathbf{M}_{R_k}, \mathbf{X}_{>k}^*)$ $= f_{A_{\leq k}}(\mathbf{X}_{k+1}^*|\mathbf{X}_{\leq k}, \mathbf{M}_{R_k}, \mathbf{X}_{>k+1}^*) f_{A_{\leq k}}(\mathbf{X}_{\leq k}, \mathbf{M}_{R_k}, \mathbf{X}_{>k+1}^*)$. Consider an identifiability assumption A_{k+1} on the distribution of $\mathbf{X}_{k+1}$ and $\mathbf{M}_{k+1}$ given $\mathbf{X}_{\leq k}, \mathbf{M}_{R_k}$ and $\mathbf{X}_{>k+1}^*$ to obtain a distribution with density $f_{A_{\leq k+1}}(\mathbf{X}_{k+1}, \mathbf{M}_{k+1}|\mathbf{X}_{\leq k}, \mathbf{M}_{R_k},$

$\mathbf{X}^*_{>k+1}$) with the NPS property $f_{A_{\leq k+1}}(\mathbf{X}^*_{k+1}|\mathbf{X}_{\leq k}, \mathbf{M}_{R_k}, \mathbf{X}^*_{>k+1}) = f_{A_{\leq k}}(\mathbf{X}^*_{k+1}|$ $\mathbf{X}_{\leq k}, \mathbf{M}_{R_k}, \mathbf{X}^*_{>k+1})$. Define

$$f_{A_{\leq k+1}}(\mathbf{X}_{\leq k+1}, \mathbf{M}_{R_k}, \mathbf{M}_{k+1}, \mathbf{X}^*_{>k+1})$$
$$\equiv f_{A_{\leq k+1}}(\mathbf{X}_{k+1}, \mathbf{M}_{k+1}|\mathbf{X}_{\leq k}, \mathbf{M}_{R_k}, \mathbf{X}^*_{>k+1}) f_{A_{\leq k}}(\mathbf{X}_{\leq k}, \mathbf{M}_{R_k}, \mathbf{X}^*_{>k+1}).$$

Step K. With assumption A_K on the distribution of $\mathbf{X}_K$ and $\mathbf{M}_K$ given $\mathbf{X}_{<K}$ and $\mathbf{M}_{R_{K-1}}$, obtain

$$f_{A_{\leq K}}(\mathbf{X}, \mathbf{M}_{R_{K-1}}, \mathbf{M}_K) \equiv f_{A_{\leq K}}(\mathbf{X}_K, \mathbf{M}_K|\mathbf{X}_{<K}, \mathbf{M}_{R_{K-1}}) f_{A_{<K}}(\mathbf{X}_{<K}, \mathbf{M}_{R_{K-1}}).$$

Obtain the implied distribution of the study variables, with density $f_{A_{\leq K}}(\mathbf{X})$, from this last equation.

Remark 1. The main characteristic of the R_k subsets is that if an index does not appear in R_{k-1}, then it cannot appear in R_k, unless it is one of $t_{k-1}+1, \ldots, t_k$. The flexibility in the choosing of these subsets gives flexibility in the setting up of the identifiability assumptions: different versions of our identification approach can be obtained by making assumptions conditioning on different subsets of the missingness indicators. As long as the R_k subsets satisfy $R_k \subseteq \{t_{k-1}+1, \ldots, t_k\} \cup R_{k-1}$, Theorem 1 guarantees that the final full-data distribution is NPS.

Remark 2. The sequence for $A_1, \ldots, A_K$ follows the order of the blocks $\mathbf{X}_1, \ldots, \mathbf{X}_K$. In many cases these blocks may not have a natural order. Different orderings of the blocks lead to different sets of assumptions, thereby leading to different final full-data distributions and implied distributions of the study variables. To clarify this point, suppose that we have three blocks of variables: $\mathbf{X}_B$, $\mathbf{X}_C$, and $\mathbf{X}_D$. When $\mathbf{X}_D$ is first in the order, A_1 concerns the distribution of $\mathbf{X}_D$ and $\mathbf{M}_D$ given $\mathbf{X}^*_B$ and $\mathbf{X}^*_C$; likewise, when $\mathbf{X}_C$ is first in the order, A_1 concerns the distribution of $\mathbf{X}_C$ and $\mathbf{M}_C$ given $\mathbf{X}^*_B$ and $\mathbf{X}^*_D$. Similarly, A_2 and A_3 also will change depending on the order of the variables, thereby implying changes in the final full-data distribution.

3.2. Non-parametric saturation

The previous presentation makes it clear that the identifying assumptions $A_1, \ldots, A_K$ allow us to identify $f_{A_{\leq K}}(\mathbf{X}, \mathbf{M}_{R_{K-1}}, \mathbf{M}_K)$, and furthermore, $f_{A_{\leq k}}(\mathbf{M}_{S_k}|$ $\mathbf{X}_{\leq k}, \mathbf{M}_{R_k}, \mathbf{X}^*_{>k})$ for each $k < K$, although each of these conditional densities remains unused after step k in the procedure. A full-data distribution with density $\tilde{f}_{A_{\leq K}}(\mathbf{X}, \mathbf{M})$ that encodes $A_1, \ldots, A_K$ can be expressed as

$$\tilde{f}_{A_{\leq K}}(\mathbf{X}, \mathbf{M}) = f_{A_{\leq K}}(\mathbf{X}, \mathbf{M}_{R_{K-1}}, \mathbf{M}_K) \tilde{f}_{A_{\leq K}}(\mathbf{M}_{S_1}, \ldots, \mathbf{M}_{S_{K-1}}|\mathbf{X}, \mathbf{M}_{R_{K-1}}, \mathbf{M}_K),$$

where the second factor can be written as $\prod_{k=1}^{K-1} \tilde{f}_{A_{\leq K}}(\mathbf{M}_{S_k}|\mathbf{X}, \mathbf{M}_{S_{>k}}, \mathbf{M}_{R_{K-1}}, \mathbf{M}_K)$, with $S_{>k} \equiv S_{k+1} \cup \cdots \cup S_{K-1}$, and $\mathbf{M}_{S_{>k}} \equiv (\mathbf{M}_{S_{k+1}}, \ldots, \mathbf{M}_{S_{K-1}})$. From the definition of the sets S_k and R_k, it is easy to see that $S_{>k} \cup R_{K-1} = R_k \cup \{t_k + 1, \ldots, t_{K-1}\}$, and therefore we can write $\tilde{f}_{A_{\leq K}}(\mathbf{M}_{S_k}|\mathbf{X}, \mathbf{M}_{S_{>k}}, \mathbf{M}_{R_{K-1}}, \mathbf{M}_K) = \tilde{f}_{A_{\leq K}}(\mathbf{M}_{S_k}|\mathbf{X}, \mathbf{M}_{R_k}, \mathbf{M}_{>k})$.

The sequential identification procedure does not identify any $\tilde{f}_{A_{\leq K}}(\mathbf{M}_{S_k}|\mathbf{X}, \mathbf{M}_{R_k}, \mathbf{M}_{>k})$, but only $f_{A_{\leq k}}(\mathbf{M}_{S_k}|\mathbf{X}_{\leq k}, \mathbf{M}_{R_k}, \mathbf{X}^*_{>k})$, that is, it identifies the distribution of $\mathbf{M}_{S_k}$ given the variables $\mathbf{X}_{\leq k}$, the missingness indicators $\mathbf{M}_{R_k}$, and the materialized variables $\mathbf{X}^*_{>k}$, but not given the missing variables among $\mathbf{X}_{>k}$ according to $\mathbf{M}_{>k}$. Nevertheless, the full specification of $\tilde{f}_{A_{\leq K}}(\mathbf{M}_{S_k}|\mathbf{X}, \mathbf{M}_{R_k}, \mathbf{M}_{>k})$ is irrelevant given that any such conditional distribution that agrees with $f_{A_{\leq k}}(\mathbf{M}_{S_k}|\mathbf{X}_{\leq k}, \mathbf{M}_{R_k}, \mathbf{X}^*_{>k})$ leads to the same $f_{A_{\leq K}}(\mathbf{X})$. One such distribution satisfies

$$\tilde{f}_{A_{\leq K}}(\mathbf{M}_{S_k}|\mathbf{X}, \mathbf{M}_{R_k}, \mathbf{M}_{>k}) = f_{A_{\leq k}}(\mathbf{M}_{S_k}|\mathbf{X}_{\leq k}, \mathbf{M}_{R_k}, \mathbf{X}^*_{>k}),\ k < K. \tag{3.1}$$

Here the conditional distribution of $\mathbf{M}_{S_k}$ given $\mathbf{X}$, $\mathbf{M}_{R_k}$, and $\mathbf{M}_{>k}$ does not depend on the missing variables among $\mathbf{X}_{>k}$ according to $\mathbf{M}_{>k}$. This guarantees the existence of a full-data distribution with density

$$\tilde{f}_{A_{\leq K}}(\mathbf{X}, \mathbf{M}) = f_{A_{\leq K}}(\mathbf{X}, \mathbf{M}_{R_{K-1}}, \mathbf{M}_K) \prod_{k=1}^{K-1} f_{A_{\leq k}}(\mathbf{M}_{S_k}|\mathbf{X}_{\leq k}, \mathbf{M}_{R_k}, \mathbf{X}^*_{>k}), \tag{3.2}$$

which encodes the assumptions $A_1, \ldots, A_K$. Theorem 1 guarantees that this construction leads to NPS full-data distributions.

Theorem 1. *Let $R_1, \ldots, R_{K-1}$ be a sequence of subsets such that $R_k \subseteq \{t_{k-1} + 1, \ldots, t_k\} \cup R_{k-1}$. Let $A_1, \ldots, A_K$ be a sequence of identifying assumptions, A_k being an assumption on the conditional distribution of $\mathbf{X}_k$ and $\mathbf{M}_k$ given $\mathbf{X}_{<k}$, $\mathbf{M}_{R_{k-1}}$, and $\mathbf{X}^*_{>k}$, such that for a given density $g(\mathbf{X}^*_k|\mathbf{X}_{<k}, \mathbf{M}_{R_{k-1}}, \mathbf{X}^*_{>k})$, one can construct a density $g_{A_k}(\mathbf{X}_k, \mathbf{M}_k|\mathbf{X}_{<k}, \mathbf{M}_{R_{k-1}}, \mathbf{X}^*_{>k})$ with the NPS property $g_{A_k}(\mathbf{X}^*_k|\mathbf{X}_{<k}, \mathbf{M}_{R_{k-1}}, \mathbf{X}^*_{>k}) = g(\mathbf{X}^*_k|\mathbf{X}_{<k}, \mathbf{M}_{R_{k-1}}, \mathbf{X}^*_{>k})$. Then, given an observed-data density $f(\mathbf{X}^*)$, there exists a full-data density $\tilde{f}_{A_{\leq K}}(\mathbf{X}, \mathbf{M})$ that encodes the assumptions $A_1, \ldots, A_K$ and satisfies the NPS property $\tilde{f}_{A_{\leq K}}(\mathbf{X}^*) = f(\mathbf{X}^*)$.*

Proof. We explained how assumptions $A_1, \ldots, A_K$ along with the extra assumption in (3.1) lead to the full-data density in (3.2). We now show the NPS property of (3.2). To start, we integrate (3.2) over the missing variables in $\mathbf{X}_K$ according to $\mathbf{M}_K$. Since none of the factors in $\prod_{k=1}^{K-1} f_{A_{\leq k}}(\mathbf{M}_{S_k}|\mathbf{X}_{\leq k}, \mathbf{M}_{R_k}, \mathbf{X}^*_{>k})$ depend on these missing variables, we obtain

$$f_{A_{\le K-1}}(\mathbf{X}_{\le K-1}, \mathbf{M}_{R_{K-1}}, \mathbf{X}^*_K) \prod_{k=1}^{K-1} f_{A_{\le k}}(\mathbf{M}_{S_k}|\mathbf{X}_{\le k}, \mathbf{M}_{R_k}, \mathbf{X}^*_{>k})$$

$$= f_{A_{\le K-1}}(\mathbf{X}_{\le K-1}, \mathbf{M}_{R_{K-1}}, \mathbf{M}_{S_{K-1}}, \mathbf{X}^*_K) \prod_{k=1}^{K-2} f_{A_{\le k}}(\mathbf{M}_{S_k}|\mathbf{X}_{\le k}, \mathbf{M}_{R_k}, \mathbf{X}^*_{>k})$$

$$= f_{A_{\le K-1}}(\mathbf{X}_{\le K-1}, \mathbf{M}_{R_{K-2}}, \mathbf{M}_{K-1}, \mathbf{X}^*_K) \prod_{k=1}^{K-2} f_{A_{\le k}}(\mathbf{M}_{S_k}|\mathbf{X}_{\le k}, \mathbf{M}_{R_k}, \mathbf{X}^*_{>k}). \quad (3.3)$$

Similarly, we now integrate (3.3) over the missing variables in $\mathbf{X}_{K-1}$ according to $\mathbf{M}_{K-1}$. Given that none of the factors in $\prod_{k=1}^{K-2} f_{A_{\le k}}(\mathbf{M}_{S_k}|\mathbf{X}_{\le k}, \mathbf{M}_{R_k}, \mathbf{X}^*_{>k})$, depend on these missing variables, and given the way $f_{A_{\le K-1}}(\mathbf{X}_{\le K-1}, \mathbf{M}_{R_{K-2}}, \mathbf{M}_{K-1}, \mathbf{X}^*_K)$ is constructed (see generic step $k+1$ in Section 3.1), we obtain

$$f_{A_{\le K-2}}(\mathbf{X}_{\le K-2}, \mathbf{M}_{R_{K-2}}, \mathbf{X}^*_{>K-2}) \prod_{k=1}^{K-2} f_{A_{\le k}}(\mathbf{M}_{S_k}|\mathbf{X}_{\le k}, \mathbf{M}_{R_k}, \mathbf{X}^*_{>k}).$$

These arguments and process can be repeated, sequentially integrating over the missing variables in $\mathbf{X}_k$ according to $\mathbf{M}_k$, $k = K-2, \ldots, 1$, finally obtaining the observed-data density $f(\mathbf{X}^*)$.

3.3. Special cases

Two special sequential identification schemes can be derived from our general presentation. One is obtained when for all k we take $R_k = \{t_{k-1}+1, \ldots, t_k\} \cup R_{k-1}$, $S_k = \emptyset$, and therefore $\mathbf{M}_{R_k} = \mathbf{M}_{\le k}$. In this case, each A_{k+1} is on the distribution of $\mathbf{X}_{k+1}$ and $\mathbf{M}_{k+1}$ given $\mathbf{X}_{\le k}, \mathbf{M}_{\le k}$ and $\mathbf{X}^*_{>k+1}$, so the assumption conditions on the whole set of missingness indicators $\mathbf{M}_{\le k}$ and not just on a subset of these. The other is obtained when for all k we take $R_k = \emptyset$, $S_k = \{t_{k-1}+1, \ldots, t_k\}$, and so $\mathbf{M}_{S_k} = \mathbf{M}_k$. In this case, each A_{k+1} is on the distribution of $\mathbf{X}_{k+1}$ and $\mathbf{M}_{k+1}$ given $\mathbf{X}_{\le k}$ and $\mathbf{X}^*_{>k+1}$, thus each assumption conditions on none of the missingness indicators $\mathbf{M}_{\le k}$.

3.4. Connection with the mechanisms of Robins (1997)

An important particular case of our sequential identification strategy is obtained when all $\mathbf{M}_{S_k} = \mathbf{M}_k$ and each A_k is taken to be a conditional MAR assumption, that is, when we assume that $f(\mathbf{M}_k = \mathbf{m}_k|\mathbf{X}_{\le k-1}, \mathbf{X}_k, \mathbf{X}^*_{>k}) = f(\mathbf{M}_k = \mathbf{m}_k|\mathbf{X}_{\le k-1}, \mathbf{X}_{k,\bar{\mathbf{m}}_k}, \mathbf{X}^*_{>k})$, with $\mathbf{X}_{k,\bar{\mathbf{m}}_k}$ being the observed variables among $\mathbf{X}_k$ according to $\mathbf{m}_k$. Along with (3.1), this leads to the combined assumption

$$f(\mathbf{M}_k = \mathbf{m}_k|\mathbf{X}, \mathbf{M}_{>k}) = f(\mathbf{M}_k = \mathbf{m}_k|\mathbf{X}_{<k}, \mathbf{X}_{k,\bar{\mathbf{m}}_k}, \mathbf{X}^*_{>k}). \quad (3.4)$$

The missingness mechanism derived from this approach corresponds to the group permutation missingness of Robins (1997). When each block contains only one variable, it corresponds to the permutation missingness mechanism of Robins (1997). If the ordering of the variables or blocks of variables is regarded as temporal, as in a longitudinal study or a survey that asks questions in a fixed sequence, Robins (1997) interpreted (3.4) as follows: the nonresponse propensity at the current time period depends on the values of study variables in the previous time periods, whether observed or not, but not on what is missing in the present and future time periods.

If the order of the blocks of variables was reversed, that is, if A_1 was on the distribution of $\mathbf{X}_K$ and $\mathbf{M}_K$ given $\mathbf{X}^*_{<K}$, A_2 was on the distribution of $\mathbf{X}_{K-1}$ and $\mathbf{M}_{K-1}$ given $\mathbf{X}^*_{<K-1}$ and $\mathbf{X}_K$, and so on, then we would have the following interpretation: the nonresponse propensity at the current time period depends on the values of study variables in the future time periods, whether observed or not, but not on what is missing in the present and past time periods. This interpretation is arguably easier to explain in the context of respondents answering a questionnaire. The nonresponse propensity for question t can depend on the respondent's answers to questions that appear later in the questionnaire and to questions that she has already answered, but not on the information that she has not revealed.

4. Applications

4.1. Sequential identification for two categorical variables

Consider two categorical random variables $X_1 \in \mathcal{X}_1 = \{1, \ldots, I\}$ and $X_2 \in \mathcal{X}_2 = \{1, \ldots, J\}$, and let M_1 and M_2 be their missingness indicators. Let $\mathbb{P}$ denote the joint distribution of (X_1, X_2, M_1, M_2). The observed-data distribution corresponds to the probabilities

$$\begin{aligned}
\pi_{ij00} &\equiv \mathbb{P}(X_1^* = i, X_2^* = j) = \mathbb{P}(X_1 = i, X_2 = j, M_1 = 0, M_2 = 0),\\
\pi_{i+01} &\equiv \mathbb{P}(X_1^* = i, X_2^* = *) = \mathbb{P}(X_1 = i, M_1 = 0, M_2 = 1),\\
\pi_{+j10} &\equiv \mathbb{P}(X_1^* = *, X_2^* = j) = \mathbb{P}(X_2 = j, M_1 = 1, M_2 = 0),\\
\pi_{++11} &\equiv \mathbb{P}(X_1^* = *, X_2^* = *) = \mathbb{P}(M_1 = 1, M_2 = 1),
\end{aligned}$$

for $i \in \mathcal{X}_1$, $j \in \mathcal{X}_2$. We seek to construct a full-data distribution $\mathbb{P}_{A_{\leq 2}}(X_1, X_2, M_1, M_2)$ from the observed-data distribution $\mathbb{P}(X_1^*, X_2^*)$ by imposing assumptions A_1 and A_2, so that $\mathbb{P}_{A_{\leq 2}}(X_1^*, X_2^*) = \mathbb{P}(X_1^*, X_2^*)$, that is, we want $\mathbb{P}_{A_{\leq 2}}$ to be NPS.

To use the general identification strategy presented in Section 3 we define

each variable as its own block. With only two variables, set R_1 can be either $R_1 = \{1\}$ or $R_1 = \emptyset$. We present examples below corresponding to these options.

Example 1. Consider $R_1 = \{1\}$, $S_1 = \emptyset$, and the identifying assumptions: $A_1: X_1 \perp\!\!\!\perp M_1 | X_2^*$; and $A_2: X_2 \perp\!\!\!\perp M_2 | M_1, X_1$.

Under A_1, $\mathbb{P}_{A_1}(X_1, M_1 | X_2^*) = \mathbb{P}_{A_1}(X_1 | X_2^*)\mathbb{P}_{A_1}(M_1 | X_2^*) = \mathbb{P}(X_1 | X_2^*, M_1 = 0)\mathbb{P}(M_1 | X_2^*)$, where $\mathbb{P}(X_1 | X_2^*, M_1 = 0)$ and $\mathbb{P}(M_1 | X_2^*)$ are identified from the observed data distribution. When $X_2^* = j \neq *$, $\mathbb{P}(X_1 = i | X_2^* = j, M_1 = 0) = \mathbb{P}(X_1 = i | X_2 = j, M_1 = 0, M_2 = 0) = \pi_{ij00}/\pi_{+j00}$, and $\mathbb{P}(M_1 = m_1 | X_2^* = j) = \mathbb{P}(M_1 = m_1 | X_2 = j, M_2 = 0) = \pi_{+jm_10}/\pi_{+j+0}$. Similarly, when $X_2^* = *$ we find $\mathbb{P}(X_1 = i | X_2^* = *, M_1 = 0) = \pi_{i+01}/\pi_{++01}$ and $\mathbb{P}(M_1 = m_1 | X_2^* = *) = \pi_{++m_11}/\pi_{+++1}$. Since $\mathbb{P}(X_2^*)$ can be obtained from the observed-data distribution as $\mathbb{P}(X_2 = j, M_2 = 0) = \pi_{+j+0}$ when $X_2^* = j \neq *$, and as $\mathbb{P}(M_2 = 1) = \pi_{+++1}$ when $X_2^* = *$, using $\mathbb{P}_{A_1}(X_1, M_1 | X_2^*)$ we obtain a joint distribution for (X_1, M_1, X_2^*) that relies on A_1, defined as $\mathbb{P}_{A_1}(X_1, M_1, X_2^*) \equiv \mathbb{P}_{A_1}(X_1, M_1 | X_2^*)\mathbb{P}(X_2^*)$. Note that $\mathbb{P}_{A_1}$ can be written as an explicit function of the observed-data distribution.

We now use $\mathbb{P}_{A_1}$ and identifying assumption A_2 to obtain $\mathbb{P}_{A_{\leq 2}}(X_2, M_2 | X_1, M_1)$. From the definition of X_2^*, $\mathbb{P}_{A_1}(X_1, M_1, X_2^*)$ can be written as $\mathbb{P}_{A_1}(X_1, M_1, X_2, M_2 = 0)$ when $X_2^* \neq *$ and $\mathbb{P}_{A_1}(X_1, M_1, M_2 = 1)$ when $X_2^* = *$. From this we can obtain

$$\mathbb{P}_{A_1}(M_2{=}1 | X_1, M_1) = \frac{\mathbb{P}_{A_1}(X_1, M_1, M_2{=}1)}{\mathbb{P}_{A_1}(X_1, M_1, M_2{=}1) + \sum_{x_2 \in \mathcal{X}_2} \mathbb{P}_{A_1}(X_1, M_1, X_2{=}x_2, M_2{=}0)},$$

$$\mathbb{P}_{A_1}(X_2 | X_1, M_1, M_2 = 0) = \frac{\mathbb{P}_{A_1}(X_1, M_1, X_2, M_2 = 0)}{\sum_{x_2 \in \mathcal{X}_2} \mathbb{P}_{A_1}(X_1, M_1, X_2 = x_2, M_2 = 0)}.$$

We then obtain $\mathbb{P}_{A_{\leq 2}}(X_2, M_2 | X_1, M_1) = \mathbb{P}_{A_{\leq 2}}(X_2 | X_1, M_1)\mathbb{P}_{A_{\leq 2}}(M_2 | X_1, M_1) = \mathbb{P}_{A_1}(X_2 | X_1, M_1, M_2 = 0)\mathbb{P}_{A_1}(M_2 | X_1, M_1)$, which gives us a way to obtain $\mathbb{P}_{A_{\leq 2}}(X_2, M_2 | X_1, M_1)$ as a function of the distribution $\mathbb{P}_{A_1}$, which in turn is a function of the observed-data distribution. The final full-data distribution is obtained as $\mathbb{P}_{A_{\leq 2}}(X_1, M_1, X_2, M_2) \equiv \mathbb{P}_{A_{\leq 2}}(X_2, M_2 | X_1, M_1)\mathbb{P}_{A_1}(X_1, M_1)$, where $\mathbb{P}_{A_1}(X_1, M_1)$ can be obtained from $\mathbb{P}_{A_1}$. After some algebra we find

$$\begin{aligned}&\mathbb{P}_{A_{\leq 2}}(X_1 = i, X_2 = j, M_1 = m_1, M_2 = m_2)\\ &= \frac{(\pi_{ij00}/\pi_{+j00})\pi_{+jm_10}}{\{\sum_l (\pi_{il00}/\pi_{+l00})\pi_{+lm_10}\}^{m_2}} \left(\frac{\pi_{i+01}}{\pi_{++01}}\pi_{++m_11}\right)^{m_2}.\end{aligned}$$

It is easy to see that $\mathbb{P}_{A_{\leq 2}}$ is NPS, that is $\mathbb{P}_{A_{\leq 2}}(X_1^*, X_2^*) = \mathbb{P}(X_1^*, X_2^*)$. From the final distribution $\mathbb{P}_{A_{\leq 2}}(X_1, X_2, M_1, M_2)$ we can now obtain

$$\mathbb{P}_{A_{\leq 2}}(X_1 = i, X_2 = j) =$$

$$\pi_{ij00} + \pi_{i+01}\frac{\pi_{ij00}}{\pi_{i+00}} + \pi_{+j10}\frac{\pi_{ij00}}{\pi_{+j00}} + \pi_{++11}\frac{\pi_{i+01}}{\pi_{++01}}\frac{(\pi_{ij00}/\pi_{+j00})\pi_{+j10}}{\sum_l(\pi_{il00}/\pi_{+l00})\pi_{+l10}}, \quad (4.1)$$

which is the distribution of inferential interest.

We stress that the final full-data distribution is not invariant to the order in which the blocks of variables appear in the sequence of assumptions. From (4.1) it is clear that the final distribution of the study variables would be different had we identified a distribution for (X_1^*, X_2, M_2) first. Indeed, if we were to follow the steps in the previous example but reversing the order of the variables, then we would be assuming that $X_2 \perp\!\!\!\perp M_2|X_1^*$ and $X_1 \perp\!\!\!\perp M_1|M_2, X_2$, which are different from A_1 and A_2 in this example.

Example 2. Consider $R_1 = \emptyset$, $S_1 = \{1\}$, and the identifying assumptions $B_1:\ X_1 \perp\!\!\!\perp M_1|X_2^*$, and $B_2:\ X_2 \perp\!\!\!\perp M_2|X_1$.

Assumption B_1 is the same as A_1 in Example 1, and so $\mathbb{P}_{B_1}(X_1, M_1, X_2^*) = \mathbb{P}_{A_1}(X_1, M_1, X_2^*)$. Assumption B_2 is made conditioning only on X_1, so we need to marginalize over M_1 to obtain $\mathbb{P}_{B_1}(X_1, X_2^*) = \mathbb{P}_{B_1}(X_1, M_1 = 0, X_2^*) + \mathbb{P}_{B_1}(X_1, M_1 = 1, X_2^*)$:

$$\mathbb{P}_{B_1}(X_1 = i, X_2^* = j) = \mathbb{P}_{B_1}(X_1 = i, X_2 = j, M_2 = 0) = \frac{\pi_{ij00}}{\pi_{+j00}}\pi_{+j+0},$$
$$\mathbb{P}_{B_1}(X_1 = i, X_2^* = *) = \mathbb{P}_{B_1}(X_1 = i, M_2 = 1) = \frac{\pi_{i+01}}{\pi_{++01}}\pi_{+++1}.$$

From this we can obtain

$$\mathbb{P}_{B_1}(M_2 = 1|X_1) = \frac{\mathbb{P}_{B_1}(X_1, M_2 = 1)}{\mathbb{P}_{B_1}(X_1, M_2 = 1) + \sum_{x_2 \in \mathcal{X}_2}\mathbb{P}_{B_1}(X_1, X_2 = x_2, M_2 = 0)},$$
$$\mathbb{P}_{B_1}(X_2|X_1, M_2 = 0) = \frac{\mathbb{P}_{B_1}(X_1, X_2, M_2 = 0)}{\sum_{x_2 \in \mathcal{X}_2}\mathbb{P}_{B_1}(X_1, X_2 = x_2, M_2 = 0)}.$$

Using assumption B_2, we obtain $\mathbb{P}_{B_{\leq 2}}(X_2, M_2|X_1) = \mathbb{P}_{B_{\leq 2}}(X_2|X_1)\mathbb{P}_{B_{\leq 2}}(M_2|X_1) = \mathbb{P}_{B_1}(X_2|X_1, M_2 = 0)\mathbb{P}_{B_1}(M_2|X_1)$. From this we obtain $\mathbb{P}_{B_{\leq 2}}(X_1, X_2, M_2) \equiv \mathbb{P}_{B_1}(X_1)\mathbb{P}_{B_{\leq 2}}(X_2, M_2|X_1)$ as

$$\mathbb{P}_{B_{\leq 2}}(X_1 = i, X_2 = j, M_2 = m_2) = \frac{(\pi_{ij00}/\pi_{+j00})\pi_{+j+0}}{\{\sum_l(\pi_{il00}/\pi_{+l00})\pi_{+l+0}\}^{m_2}}\left(\frac{\pi_{i+01}}{\pi_{++01}}\pi_{+++1}\right)^{m_2}.$$

Marginalizing over M_2, we get

$$\mathbb{P}_{B_{\leq 2}}(X_1 = i, X_2 = j) = \pi_{+j+0}\frac{\pi_{ij00}}{\pi_{+j00}} + \pi_{+++1}\frac{\pi_{i+01}}{\pi_{++01}}\frac{(\pi_{ij00}/\pi_{+j00})\pi_{+j+0}}{\sum_l(\pi_{il00}/\pi_{+l00})\pi_{+l+0}}.$$

Assumptions B_1 and B_2 are enough to identify $\mathbb{P}_{B_{\leq 2}}(X_1, X_2, M_2)$, and thereby a distribution of the study variables $\mathbb{P}_{B_{\leq 2}}(X_1, X_2)$. Although irrelevant for obtaining the distribution of the study variables, it is worth noticing that B_1 and B_2

do not allow us to fully identify $\mathbb{P}_{B_{\leq 2}}(M_1|X_1, X_2, M_2)$. From $\mathbb{P}_{B_1}(X_1, M_1, X_2^*)$ we have $\mathbb{P}_{B_1}(M_1|X_1, X_2, M_2 = 0) = \mathbb{P}(M_1|X_2, M_2 = 0)$ and $\mathbb{P}_{B_1}(M_1|X_1, M_2 = 1) = \mathbb{P}(M_1|M_2 = 1)$, but $\mathbb{P}_{B_{\leq 2}}(M_1|X_1, X_2, M_2 = 1)$ remains unidentified. A full-data distribution $\tilde{\mathbb{P}}_{B_{\leq 2}}$ becomes identified under the extra assumption $\tilde{\mathbb{P}}_{B_{\leq 2}}(M_1|X_1, X_2, M_2 = 1) = \tilde{\mathbb{P}}_{B_{\leq 2}}(M_1|X_1, M_2 = 1)$, which corresponds to the extra assumption in (3.1).

The set of assumptions that we used in this example can be summarized in terms of the missingness mechanism $\tilde{\mathbb{P}}_{B_{\leq 2}}(M_1, M_2|X_1, X_2) = \tilde{\mathbb{P}}_{B_{\leq 2}}(M_1|X_1, X_2, M_2)\tilde{\mathbb{P}}_{B_{\leq 2}}(M_2|X_1, X_2)$, where $\tilde{\mathbb{P}}_{B_{\leq 2}}(M_1|X_1, X_2, M_2 = 1) = \mathbb{P}(M_1|M_2 = 1), \tilde{\mathbb{P}}_{B_{\leq 2}}(M_1|X_1, X_2, M_2 = 0) = \mathbb{P}(M_1|X_2, M_2 = 0)$, and $\tilde{\mathbb{P}}_{B_{\leq 2}}(M_2|X_1, X_2) = \mathbb{P}_{B_1}(M_2|X_1)$. This corresponds to the permutation missingness (PM) mechanism of Robins (1997).

As in Example 1, the full-data distribution changes when we modify the order in which the blocks of variables appear in the identifying assumptions. Changing the order of the variables in this example would correspond to making the assumptions $X_2 \perp\!\!\!\perp M_2|X_1^*$ and $X_1 \perp\!\!\!\perp M_1|X_2$.

4.2. Sequential identification for partially ignorable mechanisms

Harel and Schafer (2009) introduced different notions of the missing data being *partially ignorable*. In particular, in some scenarios one may think that the missingness is ignorable for some, but not for all the variables. For example, consider a survey with two blocks of items $\mathbf{X}_S$ and $\mathbf{X}_N$ that contain responses to sensitive and non-sensitive questions, respectively. Given the nature of these variables, one might think that the missingness among the $\mathbf{X}_N$ variables could be ignored, but not among $\mathbf{X}_S$. Our sequential identification procedure can be used to guarantee identifiability under such partially ignorable mechanisms. Our goal here is to show that we can identify a NPS full-data distribution $\tilde{f}_{A_{\leq 2}}(\mathbf{X}_S, \mathbf{X}_N, \mathbf{M}_S, \mathbf{M}_N)$ with the property that the missingness mechanism for $\mathbf{X}_N$ is partially MAR given $\mathbf{M}_S$ (Harel and Schafer (2009)), that is,

$$\tilde{f}_{A_{\leq 2}}(\mathbf{M}_N = \mathbf{m}_N|\mathbf{X}_N, \mathbf{X}_S, \mathbf{M}_S) = \tilde{f}_{A_{\leq 2}}(\mathbf{M}_N = \mathbf{m}_N|\mathbf{X}_{N,\bar{\mathbf{m}}_N},\ \mathbf{X}_S^*), \tag{4.2}$$

while $\tilde{f}_{A_{\leq 2}}(\mathbf{M}_S|\mathbf{X}_N, \mathbf{X}_S)$ is determined by some nonignorable assumption. As before, we consider $f(\mathbf{X}_S^*, \mathbf{X}_N^*)$ to be known.

Following our sequential identification procedure, we first consider an identifying assumption for the distribution of $\mathbf{X}_N$ and $\mathbf{M}_N$ given $\mathbf{X}_S^*$. We use the conditional MAR assumption:

$$A_1 :\ f(\mathbf{M}_N = \mathbf{m}_N|\mathbf{X}_N, \mathbf{X}_S^*) = f(\mathbf{M}_N = \mathbf{m}_N|\mathbf{X}_{N,\bar{\mathbf{m}}_N}, \mathbf{X}_S^*). \tag{4.3}$$

This assumption guarantees the existence of a distribution of the variables $\mathbf{X}_N$, $\mathbf{M}_N$, and $\mathbf{X}_S^*$ with density $f_{A_1}(\mathbf{X}_N, \mathbf{M}_N, \mathbf{X}_S^*) \equiv f_{A_1}(\mathbf{M}_N|\mathbf{X}_N, \mathbf{X}_S^*) f_{A_1}(\mathbf{X}_N|\mathbf{X}_S^*)$ $f(\mathbf{X}_S^*)$, where $f_{A_1}(\mathbf{X}_N|\mathbf{X}_S^*)$ can be obtained from $f(\mathbf{X}_N^*|\mathbf{X}_S^*)$ as described in page 28 of Robins (1997) and $f_{A_1}(\mathbf{M}_N|\mathbf{X}_N, \mathbf{X}_S^*)$ as in (4.3).

Taking $R_1 = \emptyset$ in our identification procedure, we can now consider any identifying assumption, say A_2, for the distribution of $\mathbf{X}_S$ and $\mathbf{M}_S$ given $\mathbf{X}_N$ that allows us to obtain $f_{A_{\leq 2}}(\mathbf{X}_S, \mathbf{M}_S|\mathbf{X}_N)$ with the NPS property $f_{A_{\leq 2}}(\mathbf{X}_S^*|\mathbf{X}_N) = f_{A_1}(\mathbf{X}_S^*|\mathbf{X}_N)$. For example, A_2 could come from one of the approaches mentioned in Section 2.2. We then define $f_{A_{\leq 2}}(\mathbf{X}_N, \mathbf{X}_S, \mathbf{M}_S) \equiv f_{A_{\leq 2}}(\mathbf{X}_S, \mathbf{M}_S|\mathbf{X}_N) f_{A_1}(\mathbf{X}_N)$.

To fully identify a full-data distribution $\tilde{f}_{A_{\leq 2}}(\mathbf{X}_S, \mathbf{X}_N, \mathbf{M}_S, \mathbf{M}_N)$ that encodes assumptions A_1 and A_2, we further require the conditional missingness mechanism $\tilde{f}_{A_{\leq 2}}(\mathbf{M}_N|\mathbf{X}_N, \mathbf{X}_S, \mathbf{M}_S)$. Under the extra assumption

$$\tilde{f}_{A_{\leq 2}}(\mathbf{M}_N|\mathbf{X}_N, \mathbf{X}_S, \mathbf{M}_S) = \tilde{f}_{A_{\leq 2}}(\mathbf{M}_N|\mathbf{X}_N, \mathbf{X}_S^*), \tag{4.4}$$

and then using A_1, we have identified a full-data distribution with the NPS property guaranteed by Theorem 1.

A possibility for the A_2 assumption could come from the itemwise conditionally independent nonresponse (ICIN) mechanism of Sadinle and Reiter (2017), which is NPS. With $\mathbf{X}_S = (X_{S1}, \ldots, X_{Sp_S})$, the ICIN assumption for $\mathbf{X}_S$ and $\mathbf{M}_S$ given $\mathbf{X}_N$ can be written as

$$X_{Sj} \perp\!\!\!\perp M_{Sj}|\mathbf{X}_{S(-j)}, \mathbf{M}_{S(-j)}, \mathbf{X}_N;\ j = 1, \ldots, p_S, \tag{4.5}$$

where $\mathbf{X}_{S(-j)}$ is the vector obtained from removing the jth entry of $\mathbf{X}_S$, likewise for $\mathbf{M}_{S(-j)}$. Our sequential identification procedure guarantees that assumptions in (4.2) and (4.5) jointly identify a NPS full-data distribution.

Example 3. For simplicity, consider $\mathbf{X}_N = X_1$ and $\mathbf{X}_S = (X_2, X_3)$. The observed-data density can be written as the product of the density of the observed variables given each missingness pattern times the probability of the missingness pattern, that is $f(\mathbf{X}_{\bar{\mathbf{m}}}, \mathbf{M} = \mathbf{m}) = f_{\mathbf{m}}(\mathbf{X}_{\bar{\mathbf{m}}})\pi_{\mathbf{m}}$, which for three variables is given by $f_{000}(X_1, X_2, X_3)\pi_{000}$, $f_{100}(X_2, X_3)\pi_{100}$, $f_{010}(X_1, X_3)\pi_{010}, \ldots, f_{011}(X_1)\pi_{011}$, and π_{111}. Assumption A_1 in (4.3) in this case is $A_1 :\ X_1 \perp\!\!\!\perp M_1|X_2^*, X_3^*$, which for all $x_2 \in \mathcal{X}_2$ and $x_3 \in \mathcal{X}_3$ can be expanded as $X_1 \perp\!\!\!\perp M_1|M_2 = 0, M_3 = 0, X_2 = x_2, X_3 = x_3$; $X_1 \perp\!\!\!\perp M_1|M_2 = 1, M_3 = 0, X_3 = x_3$; $X_1 \perp\!\!\!\perp M_1|M_2 = 0, M_3 = 1, X_2 = x_2$; and $X_1 \perp\!\!\!\perp M_1|M_2 = 1, M_3 = 1$. Using A_1 and the observed-data distribution we obtain $f_{A_1}(X_1, M_1|X_2^*, X_3^*) = f_{A_1}(X_1|X_2^*, X_3^*) f_{A_1}(M_1|X_2^*, X_3^*) = f(X_1|M_1 = 0, X_2^*, X_3^*) f(M_1|X_2^*, X_3^*)$, where $f(X_1|M_1 = 0, X_2^*, X_3^*)$ is obtained from

$$f(X_1|M_1=0,M_2=0,M_3=0,X_2,X_3)=\frac{f_{000}(X_1,X_2,X_3)}{f_{000}(X_2,X_3)},$$
$$f(X_1|M_1=0,M_2=1,M_3=0,X_3)=\frac{f_{010}(X_1,X_3)}{f_{010}(X_3)},$$
$$f(X_1|M_1=0,M_2=0,M_3=1,X_2)=\frac{f_{001}(X_1,X_2)}{f_{001}(X_2)},$$
$$f(X_1|M_1=0,M_2=1,M_3=1)=f_{011}(X_1);\text{ and } f(M_1|X_2^*,X_3^*) \text{ from}$$
$$f(M_1{=}m_1|M_2=M_3=0,X_2,X_3)\propto[f_{000}(X_2,X_3)\pi_{000}]^{I(m_1=0)}$$
$$[f_{100}(X_2,X_3)\pi_{100}]^{I(m_1=1)},$$
$$f(M_1=m_1|M_2=1,M_3=0,X_3)\propto[f_{010}(X_3)\pi_{010}]^{I(m_1=0)}[f_{110}(X_3)\pi_{110}]^{I(m_1=1)},$$
$$f(M_1=m_1|M_2=0,M_3=1,X_2)\propto[f_{001}(X_2)\pi_{001}]^{I(m_1=0)}[f_{101}(X_2)\pi_{101}]^{I(m_1=1)},$$
$$f(M_1=m_1|M_2=1,M_3=1)\propto[\pi_{011}]^{I(m_1=0)}[\pi_{111}]^{I(m_1=1)}.$$

From this we can define $f_{A_1}(X_1,M_1,X_2^*,X_3^*)\equiv f_{A_1}(X_1,M_1|X_2^*,X_3^*)f(X_2^*,X_3^*)$, where $f(X_2^*,X_3^*)$ is obtained from $f(M_2=0,M_3=0,X_2,X_3)=f_{000}(X_2,X_3)\pi_{000}$ $+f_{100}(X_2,X_3)\pi_{100}$, $f(M_2=1,M_3=0,X_3)=f_{010}(X_3)\pi_{010}+f_{110}(X_3)\pi_{110}$, $f(M_2=0,M_3=1,X_2)=f_{001}(X_2)\pi_{001}+f_{101}(X_2)\pi_{101}$, and $f(M_2=1,M_3=1)=\pi_{011}+\pi_{111}$.

We now incorporate the ICIN assumption for the distribution of (X_2,X_3,M_2,M_3) given X_1. We have $A_2:\{X_2 \perp\!\!\!\perp M_2|X_3,M_3,X_1;$ and $X_3 \perp\!\!\!\perp M_3|X_2,M_2,X_1\}$. The identification results of Sadinle and Reiter (2017) guarantee that assumption A_2 leads to a conditional distribution $f_{A_{\leq 2}}(X_2,X_3,M_2,M_3|X_1)$ with the NPS property $f_{A_{\leq 2}}(X_2^*,X_3^*|X_1)=f_{A_1}(X_2^*,X_3^*|X_1)$, where $f_{A_1}(X_2^*,X_3^*|X_1)$ can be obtained easily from $f_{A_1}(X_1,M_1,X_2^*,X_3^*)$. Section 5.1 of Sadinle and Reiter (2017) provides explicit formulae for the full-data distribution under the ICIN assumption as a function of the observed-data distribution, in the case of two variables. We can use those formulae here with $f_{A_1}(X_2^*,X_3^*|X_1)$ to obtain conditional ICIN full-data distributions that depend on X_1. To simplify the notation below we replace f_{A_1} by g, and $f_{A_{\leq 2}}$ by h, and we denote $g_{m_2m_3}(X_2,X_3|X_1)\equiv g(X_2,X_3|X_1,M_2=m_2,M_3=m_3)$ and $h_{m_2m_3}(X_2,X_3|X_1)\equiv h(X_2,X_3|X_1,M_2=m_2,M_3=m_3)$. Following the formulae of Sadinle and Reiter (2017), we obtain $h_{00}(X_2,X_3|X_1)=g_{00}(X_2,X_3|X_1), h_{01}(X_2,X_3|X_1)=g_{00}(X_2,X_3|X_1)g_{01}(X_2|X_1)/g_{00}(X_2|X_1)$, $h_{10}(X_2,X_3|X_1)=g_{00}(X_2,X_3|X_1)g_{10}(X_3|X_1)/g_{00}(X_3|X_1)$,

$$h_{11}(X_2,X_3|X_1)\propto\frac{g_{00}(X_2,X_3|X_1)g_{01}(X_2|X_1)g_{10}(X_3|X_1)}{g_{00}(X_2|X_1)g_{00}(X_3|X_1)},$$

and $h(M_2=m_2,M_3=m_3|X_1)=g(M_2=m_2,M_3=m_3|X_1)$.

In all,

$$f_{A_{\leq 2}}(X_1, X_2, X_3, M_2 = m_2, M_3 = m_3) = \\ h_{m_2 m_3}(X_2, X_3|X_1) g(M_2 = m_2, M_3 = m_3, X_1),$$

from which we can obtain the distribution of the study variables $f_{A_{\leq 2}}(X_1, X_2, X_3)$. A full-data density $\tilde{f}_{A_{\leq 2}}(X_1, X_2, X_3, M_1, M_2, M_3)$ becomes identified under the extra assumption (4.4). This distribution therefore encodes the partial ignorability assumption (4.2) for the missingness in X_1 and the ICIN assumption (4.5) for (X_2, X_3, M_2, M_3) given X_1.

4.3. Usage in sensitivity analysis

To illustrate how this approach can be used for sensitivity analysis, we use data related to the 1991 plebiscite in which Slovenians voted for independence from Yugoslavia (Rubin, Stern and Vehovar (1995)). The data come from the Slovenian public opinion survey, which contained the questions: X_I: are you in favor of Slovenia's independence? X_S: are you in favor of Slovenia's secession from Yugoslavia? X_A: will you attend the plebiscite? We call these the Independence, Secession, and Attendance questions, respectively. The possible responses to each of these were YES, NO, and DON'T KNOW. We follow Rubin, Stern and Vehovar (1995) in treating DON'T KNOW as missing data.

We use the missingness mechanism of Example 3, and compare it with an ignorable approach, a pattern mixture model (PMM) under the complete-case missing-variable restriction (Little (1993)), and the ICIN approach of Sadinle and Reiter (2017) that assumes $X_j \perp\!\!\!\perp M_j|\mathbf{X}_{-j}, \mathbf{M}_{-j};\ j = I, S, A$. The Attendance question is arguably the least sensitive of the three questions, so it seems reasonable to consider a partially ignorable mechanism where the nonresponse for X_A is ignorable given M_I and M_S, as in (4.2), and the nonresponse for X_I and X_S satisfy the ICIN assumption conditioning on X_A, as in (4.5) in Example 3. Nothing prevents us from using this approach exchanging the roles of the variables, so we also consider two other partially ignorable missingness mechanisms, depending on whether we take the nonresponse for X_I or for X_S as ignorable.

To implement these approaches, we first use a Bayesian approach to estimate the observed-data distribution. The observed data can be organized in a three-way contingency table with cells corresponding to each element of $\{$YES, NO, DON'T KNOW$\}^3$, as presented in Rubin, Stern and Vehovar (1995). Treating these data as a random sample from a multinomial distribution, we take a conjugate prior distribution for the cell probabilities: symmetric Dirichlet with parameter

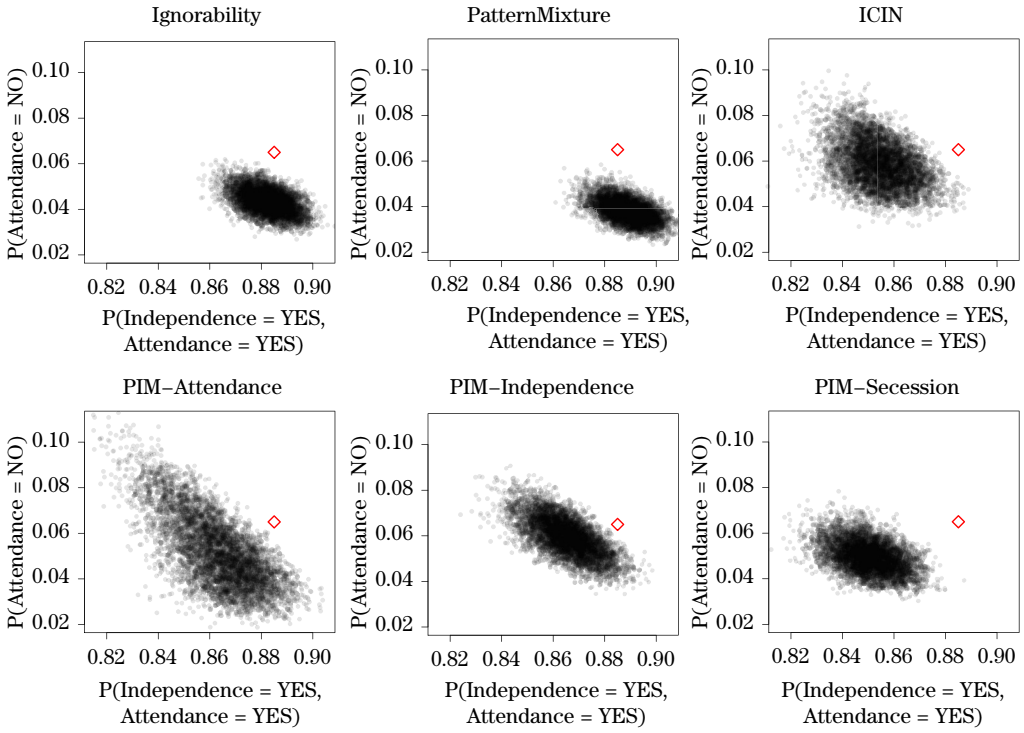

Figure 2. Samples from joint posterior distributions of $\mathbb{P}$(Independence = Yes, Attendance = Yes) and $\mathbb{P}$(Attendance = No). Pattern mixture model under the complete-case missing-variable restriction. The three partially ignorable missingness (PIM) mechanisms correspond to which variable we take as having ignorable missingness. The plebiscite results are represented by ◇. These are shown to illustrate differences between approaches and not to declare better vs worse assumptions for these data.

1/27. We took 5,000 draws from the posterior distribution of the observed-data distribution, and for each of these we applied the formulae presented in Example 3 to obtain posterior draws of the full-data distribution under each of the three partially ignorable mechanisms. We used a similar approach to obtain posterior draws of the full-data distribution under ICIN, PMM, and ignorability. For each of the approaches we then obtained draws of the implied probabilities for the items.

Figure 2 displays 5,000 draws from the joint posterior distribution of $\mathbb{P}$(Independence = YES, Attendance = YES) and $\mathbb{P}$(Attendance = NO) under each of the six missingness mechanisms considered. Despite the fact that all of these approaches agree in their fit to the observed data, we obtain quite different inferences. When inferences are so sensitive to the identifying assumptions, perhaps

the most honest way to proceed is to report all the results under all assumptions deemed plausible given the context.

5. Discussion

The sequential identification procedure can be set up in many different ways, leading to different possibilities for constructing nonignorable missingness mechanisms. The main differences among these possibilities lie in the assumptions about how missingness from any one block of variables affects missingness in other blocks, as illustrated in the examples of Section 3.3 and Section 4. In general, the procedure allows for different levels of dependence on missing variables while ensuring non-parametric saturated models, which provides a useful framework for sensitivity analysis.

Although we considered K blocks of variables, we expect that in practice most analysts would use $K = 2$ blocks when the variables do not naturally fall into ordered blocks of variables. For example, analysts may want to partition variables into one group that requires careful assessment of sensitivity to various missingness mechanisms, such as outcome variables in regression modeling with high fractions of missingness, and a second group that can be treated with generic missingness mechanisms like conditional MAR, such as covariates with low fractions of missingness. These cases require partially ignorable mechanisms like those in Section 4.2. Another scenario arises when analysts have prior information on how the missingness occurs for a set of variables, but not for the rest. As well, analysts might have auxiliary information on the marginal distribution of a few variables, perhaps from a census or other surveys, that enable the identification of mechanisms where the probability of nonresponse for a variable depends explicitly on the variable itself (Hirano et al. (2001); Deng et al. (2013)).

Our sequential identification procedure provides a constructive way of obtaining estimated full-data distributions from estimated observed-data distributions while ensuring non-parametric saturated models. To implement these approaches, one needs sufficient numbers of observations for each missing data pattern, so as to allow accurate non-parametric estimation of the observed-data distribution. This can be challenging in modest-sized samples with large numbers of variables. Of course, this is the case with most methods for handling missing data, including pattern mixture models. In such cases, one may have to sacrifice non-parametric saturated modeling of the observed data in favor of parametric models.

Acknowledgment

This research was supported by the grant NSF SES 11-31897.

References

Daniels, M. J. and Hogan, J. W. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Chapman and Hall/CRC, Boca Raton, Florida.

Deng, Y., Hillygus, D. S., Reiter, J. P., Si, Y. and Zheng, S. (2013). Handling attrition in longitudinal studies: the case for refreshment samples. *Statist. Sci.* **28**, 238–256.

Gill, R. D., van der Laan, M. J. and Robins, J. M. (1997). Coarsening at random: characterizations, conjectures, counter-examples. *In Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis* (Edited by Lin, D. Y. and Fleming, T. R.), 255–294.

Harel, O. and Schafer, J. L. (2009). Partial and latent ignorability in missing-data problems. *Biometrika* **96**, 37–50.

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

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *J. Amer. Statist. Assoc.* **88**, 125–134.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. 2nd edition, Wiley, Hoboken, New Jersey.

Robins, J. M. (1997). Non-response models for the analysis of non-monotone non-ignorable missing data. *Stat. Med.* **16**, 21–37.

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

Rubin, D. B., Stern, H. S. and Vehovar, V. (1995). Handling "Don't know" survey responses: The case of the Slovenian plebiscite *J. Amer. Statist. Assoc.* **90**, 822–828.

Sadinle, M. and Reiter, J. P. (2017). Itemwise conditionally independent nonresponse modeling for incomplete multivariate data. *Biometrika* **104**, 207–220.

Tchetgen Tchetgen, E. J., Wang, L. and Sun, B. (2016). Discrete choice models for nonmonotone nonignorable missing data: identification and inference. ArXiv: 1607.02631

Vansteelandt, S., Goetghebeur, E., Kenward, M. G. and Molenberghs, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statist. Sinica* **16**, 953–979.

Zhou, Y., Little, R. J. A. and Kalbfleisch, J. D. (2010). Block-conditional missing at random models for missing data. *Statist. Sci.* **25**, 517–532.

Department of Biostatistics, University of Washington, Seattle, WA 98195, U.S.A.

E-mail: msadinle@uw.edu

Department of Statistical Science, Box 90251, Duke University, Durham, NC 27708, U.S.A.

E-mail: jerry@stat.duke.edu

(Received June 2016; accepted February 2017)