Sequentially additive nonignorable missing data modeling using auxiliary marginal information

Mauricio Sadinle and Jerome P. Reiter

University of Washington and Duke University

Abstract

We study a class of missingness mechanisms, called sequentially additive nonignorable, for modeling multivariate data with item nonresponse. These mechanisms explicitly allow the probability of nonresponse for each variable to depend on the value of that variable, thereby representing nonignorable missingness mechanisms. These missing data models are identified by making use of auxiliary information on marginal distributions, such as marginal probabilities for multivariate categorical variables or moments for numeric variables. We present theory proving identification results, and illustrate the use of these mechanisms in an application.

Keywords: Information projection; Missing not at random; Nonmonotone nonresponse; Nonparametric identification; Observational equivalence.

1 Introduction

When data values are unintentionally missing, analysts generally cannot estimate the true probabilistic mechanism generating the missingness from the observed data alone. To proceed with statistical inference, they have to make unverifiable assumptions on how the missingness arises. These identifying assumptions correspond to restrictions on the joint distribution of the study variables and their missingness indicators. Often in practice, analysts specify such restrictions on the conditional distribution of the missingness indicators given the study variables; for example, they assume the data are missing at random (Rubin, 1976). In many contexts, it is desirable to use restrictions that encode nonignorable missingness mechanisms (Greenlees et al., 1982; Robins, 1997; Sadinle and Reiter, 2017). However, for multivariate data with arbitrary patterns of nonresponse, it can be difficult to

specify such nonignorable mechanisms in ways that lead to provably identifiable models (Ibrahim et al., 1999; Daniels and Hogan, 2008, Section 5.9.1).

In this article, we study a class of such missingness mechanisms, called sequentially additive nonignorable, for handling item nonresponse in multiple variables. We adopt this terminology from Hoonhout and Ridder (2018), who introduced the sequentially additive nonignorable attrition mechanism for longitudinal studies with monotone nonresponse, that is, when participants who drop out no longer return to the study. Similarly as with the attrition mechanism, the missingness mechanism explicitly allows item nonresponse in multiple variables to depend on the values of those variables. We prove that identifiability is attainable using different types of auxiliary marginal information, that is, external information on features of the distribution of the study variables. Examples of external data sources include censuses and administrative databases, which readily provide the population distribution of some variables; large surveys, which provide high-quality estimates of some population characteristics; and refreshment samples in longitudinal studies, where at later waves additional observations are drawn from the population and whose responses are fully recorded (Hirano et al., 1998, 2001; Deng et al., 2013; Hoonhout and Ridder, 2018).

The sequentially additive nonignorable missingness mechanism builds on the additive nonignorable mechanism of Hirano et al. (1998, 2001) and on the attrition mechanism of Hoonhout and Ridder (2018), with three key differences. First and most critically, Hirano et al. (1998, 2001) focused on the case of one variable subject to missingness, and Hoonhout and Ridder (2018) on multiple variables subject to a monotone missingness pattern, whereas we consider the more general case of multiple variables under nonmonotone nonresponse. Second, their identification results apply to either exclusively categorical or exclusively continuous variables, whereas our results apply to general variable types in arbitrary probability spaces. Third, our identification results can be used with different types of auxiliary marginal information, such as moments of random variables, or univariate and multivariate marginal distributions, unlike Hirano et al. (1998, 2001) and Hoonhout and Ridder (2018) which require univariate marginal distributions.

2 Set-up and preliminaries

2.1 Notation

Let the random vectors $Y = (Y_1, \ldots, Y_p)$ and $X = (X_1, \ldots, X_q)$ represent the study variables, where auxiliary marginal information on the distribution of Y is available. We formally define auxiliary marginal information in Section 2.3. In general, the variables in both X and Y might be subject to missingness. Let $M = (M_1, \ldots, M_p)$ be the vector of missingness indicators for Y, where $M_j = 1$ when Y_j is missing and $M_j = 0$ when Y_j is observed. The vector $W = (W_1, \ldots, W_q)$ of missingness indicators for X is defined analogously. We denote $Y_{<j} = (Y_1, \ldots, Y_{j-1})$, $Y_{\leq j} = (Y_1, \ldots, Y_j)$, $Y_{>j} = (Y_{j+1}, \ldots, Y_p)$, $Y_{\geq j} = (Y_j, \ldots, Y_p)$, and similarly for subvectors of X, W and M. We denote generic possible values of Y and $Y_{<j}$ by y and $y_{<j}$, respectively, and similarly for other random vectors.

Let μ_j and ν_l be dominating measures for the marginal distributions of Y_j and X_l, respectively, and let $\mu = \otimes_{j=1}^{p} \mu_j$ and $\nu = \otimes_{l=1}^{q} \nu_l$. The full-data distribution is the joint distribution of (X, Y, W, M), and we refer to its density $f(x, y, w, m)$ with respect to the product of ν, μ, and the counting measure on $\{0,1\}^{p+q}$ as the full-data density. The conditional distribution with density $f(w, m \mid x, y)$ is referred to as the missingness mechanism (Daniels and Hogan, 2008, p. 90). For 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 density of the marginal distribution of Y by $f(y)$. The set of functions u such that $\int |u(y)| f(y) \mu(dy) < \infty$ is denoted $L_1\{f(y)\}$, and similarly for other densities. The linear span of a set of functions $\mathcal{U}$ is denoted $\langle \mathcal{U} \rangle$, and its closure $\overline{\langle \mathcal{U} \rangle}$.

A missingness pattern for the Y variables is represented by $m = (m_1, \ldots, m_p) \in \{0,1\}^p$. Given m we define $\bar{m} = 1_p - m$ to be the indicator vector of observed Y variables, where 1_p is a vector of ones of length p. We define $Y_m = (Y_j : m_j = 1)$ to be the missing Y variables and $Y_{\bar{m}} = (Y_j : \bar{m}_j = 1)$ to be the observed Y variables according to m. We define $w = (w_1, \ldots, w_q) \in \{0,1\}^q$, $\bar{w}$, X_w and $X_{\bar{w}}$ in an analogous fashion. The observed-data

distribution is the distribution involving the missingness indicators and the corresponding observed variables, with density

$$f(x_{\bar{w}}, y_{\bar{m}}, w, m) = \iint f(x, y, w, m)\nu(dx_w)\mu(dy_m). \tag{1}$$

A more compact way of representing the observed-data distribution is obtained by introducing what Sadinle and Reiter (2018) call materialized variables, defined as

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

where $*$ is simply a placeholder for missingness (Rubin, 1976). We define analogously X_l^*, and denote $Y^* = (Y_1^*, \ldots, Y_p^*)$ and $X^* = (X_1^*, \ldots, X_q^*)$. All the observable information from the sampling process can be obtained from the materialized variables: if $Y_j^* = *$ then Y_j is not observed, and if $Y_j^* = y_j \neq *$ then Y_j is observed and $Y_j = y_j$. This means that the distribution of (X^*, Y^*) is nothing but a different way of representing the observed-data distribution. Therefore, with some abuse of notation, the observed-data density can be written in terms of X^* and Y^*, that is, $f(x^*, y^*) \equiv f(x_{\bar{w}}, y_{\bar{m}}, w, m)$, where $y^* = (y_1^*, \ldots, y_p^*)$ with $y_j^* = *$ if $m_j = 1$ and $y_j^* = y_j$ if $m_j = 0$, likewise for x^*. We also often use a hybrid notation as in $f(x, y^*) \equiv f(x, y_{\bar{m}}, m)$.

For our identification results we will work with generic random vectors (X^*, Y^*), but for estimation we will assume to have observations that are drawn as independent and identically distributed copies of (X^*, Y^*), as explained in Section 4.3.

2.2 Identifiability

The full-data distribution cannot be identified from observed data alone in a nonparametric manner; even if we had the ability to sample indefinitely we would only be able to recover the observed-data distribution. This fact leads us to introduce some necessary definitions. We first present the concept of observational equivalence, which we borrow from Koopmans (1949).

Definition 1 (Observational equivalence). *Two full-data distributions are said to be observationally equivalent if their implied observed-data distributions are the same.*

Consider now a class of full-data distributions $\mathcal{C}_\Theta$ indexed by the parameter space Θ which is either finite- or infinite-dimensional. If we were able to observe the values of the study variables regardless of the values of their missingness indicators, we would still have to guarantee that $\mathcal{C}_\Theta$ is identifiable in the usual sense (e.g., Lehmann and Casella, 1998, p. 24).

Definition 2 (Full-data identifiability). *A class of full-data distributions $\mathcal{C}_\Theta$ is said to be full-data identifiable if there exists a bijection from Θ to $\mathcal{C}_\Theta$.*

Full-data identifiability is an elementary requirement of $\mathcal{C}_\Theta$ which simply says that the class is properly parameterized, and therefore throughout this article we will assume that this holds. Let now $\text{obs}(\mathcal{C}_\Theta)$ denote the class of observed-data distributions implied by $\mathcal{C}_\Theta$ according to (1).

Definition 3 (Identifiability). *A class of full-data distributions $\mathcal{C}_\Theta$ is said to be identifiable if there exist bijections from Θ to $\mathcal{C}_\Theta$ and from obs$(\mathcal{C}_\Theta)$ to $\mathcal{C}_\Theta$.*

The first bijection in this definition corresponds to full-data identifiability for $\mathcal{C}_\Theta$, and the second one simply tells us that we need a unique way to go back and forth from $\text{obs}(\mathcal{C}_\Theta)$ to $\mathcal{C}_\Theta$. These two bijections imply a third one between $\text{obs}(\mathcal{C}_\Theta)$ and Θ, which corresponds to the common notion of identifiability applied to $\text{obs}(\mathcal{C}_\Theta)$.

If the class of observed-data distributions $\text{obs}(\mathcal{C}_\Theta)$ is a proper subset of all observed-data distributions, then the model $\mathcal{C}_\Theta$ imposes parametric restrictions on what could be nonparametrically recovered from observed data alone. Thus, we also make use of a stricter property for a class of full-data distributions, namely, nonparametric identification, also known as nonparametric saturation or just-identification (Robins, 1997; Vansteelandt et al., 2006; Daniels and Hogan, 2008; Hoonhout and Ridder, 2018).

Definition 4 (Nonparametric identifiability). *A class of full-data distributions $\mathcal{C}_\Theta$ is said to be nonparametrically identifiable if it is identifiable and obs$(\mathcal{C}_\Theta)$ equals the set of all observed-data distributions.*

Given the bijective mapping between Θ and $\text{obs}(\mathcal{C}_\Theta)$ obtained from the identifiability requirement, we can think of a nonparametrically identifiable class as being indexed by the set of all observed-data distributions.

Two nonparametrically identifiable classes necessarily lead to full-data distributions that are observationally equivalent, and therefore the assumptions used to build such classes cannot be refuted from observed data alone. When two missing data models are observationally equivalent, any discrepancies in inferences are due entirely to the difference in the restrictions on the unidentifiable parts of the full-data distribution. Nonparametric identification additionally guarantees that these restrictions do not constrain the observed-data distribution. Nonparametric identification is therefore a basic desirable property, particularly useful for comparing inferences under different missing data assumptions.

2.3 Auxiliary marginal information

To guarantee the nonparametric identifiability of the sequentially additive nonignorable missingness mechanisms that we will introduce, we need to have access to external information on features of the distribution of some of the study variables, here represented by $Y = (Y_1, \ldots, Y_p)$. Suitable examples include the joint distribution of Y, joint marginal distributions for subsets of Y, the individual marginal distributions of each Y_j, or expected values of functions of the Y variables, including means, variances, and general moments. We provide a formal definition that encompasses all of these cases.

Definition 5 (Auxiliary marginal information). *Let the set of functions $\mathcal{U} \subseteq L_1\{f(y)\}$ contain for each $j = 1, \ldots, p$ at least one almost-surely non-constant function of y_j. If we know the value of $E[u(Y)] = \int u(y)f(y)\mu(dy) < \infty$ for each $u \in \mathcal{U}$, we refer to $\{E[u(Y)]\}_{u\in\mathcal{U}}$ as auxiliary marginal information on the distribution of $Y = (Y_1, \ldots, Y_p)$.*

For all of our results, to separate identification from estimation issues, having auxiliary marginal information on a set of variables is conceptualized as knowing the expected value of a set of functions $\mathcal{U}$ with respect to the distribution of Y. Data-based implementations are discussed in Section 4.

Definition 5 includes different types of external information. Important cases include the marginal distribution of Y, where $\mathcal{U}$ would be taken as all the integrable functions with respect to the distribution of Y; the marginal distributions of each Y_j, where $\mathcal{U}$ would be taken as all the integrable functions with respect to the distributions of each Y_j seen as functions of the whole vector y; the means of the variables, where $\mathcal{U}$ would contain p functions $u_1, \ldots, u_p$ where $u_j(y) = y_j$; general moments of the variables, where $\mathcal{U}$ would contain functions $u(y) = y_j^z$ for $z \neq 0$; among many others. Intuitively, the richer the set $\mathcal{U}$ the more flexible the class of missingness mechanisms that we will be able to identify.

Auxiliary data on marginal distributions can be available in many contexts and in many forms. For example, published data products from national censuses of businesses, farms, and people provide marginal distributions for many variables commonly included in sample surveys of those populations, such as number of employees and demographic characteristics. Indeed, sample surveys routinely use known margins from censuses and sampling frames in weight calibration and generalized regression estimation (e.g., Lohr, 2010, ch. 11). Administrative databases, such as tax records, voter registration files, state education records on school children, and medical records from insurers or government programs, like Medicare in the U.S.A., can provide margins on potentially relevant populations. Large nationally representative surveys like the American Community Survey in the U.S.A. can provide estimates of marginal means and totals with small enough standard errors to be treated as essentially known. Marginal distributions of biomarker measurements may be available from calibration studies done by assay producers or government agencies. Other examples of data analyses that use auxiliary information can be found in, for example, Berrocal et al. (2013), Chatterjee et al. (2016), Guo et al. (2012), and National Academies of Sciences, Engineering, and Medicine (2017).

Naturally, analysts should carefully consider their choice of external data sources. In particular, analysts should ensure that the information is contemporaneous with, covers the same target population as, and uses the same variable definitions as the data being analyzed. Analysts should feel comfortable that the information does not contain substantial biases,

for example, from convenience sampling, measurement error, and nonresponse bias within the data used to obtain the auxiliary marginal information.

3 Sequential additive nonignorability

3.1 Extending results for univariate nonresponse

We begin with a single random variable Y subject to missingness; here, M denotes its missingness indicator, and the vector of variables X is fully observed. This is the set-up studied by Hirano et al. (1998, 2001) in the context of a longitudinal study, where the X variables are recorded at a given time point, and Y denotes a follow-up measurement that is sometimes missing due to attrition. In that context the auxiliary marginal information comes from a refreshment sample, seen as a random sample from the marginal distribution of Y, which they conceptualize as knowing $f(y)$.

Given $f(y)$, Hirano et al. (1998, 2001) showed that there exist identifiable missingness mechanisms with the form

$$\lambda[f(M = 1 \mid x, y)] = \alpha(x) + \beta(y), \tag{2}$$

for a link function λ, and for some functions α and β which are essentially unrestricted except for some integrability conditions. The model is additive in $\alpha(x)$ and $\beta(y)$ because there is not enough information to identify interactions between X and Y (Hirano et al., 1998, 2001). Examples of analyses that make use of additive nonignorable missingness mechanisms include the work in Nevo (2003), Bhattacharya (2008), and Si et al. (2015), among others.

The missingness mechanism in (2) is appealing, as it includes as special cases a missing always at random mechanism (Mealli and Rubin, 2015) when $\beta(y) = 0$ and the often-used selection model of Hausman and Wise (1979) when $\alpha(x) = 0$. Thus, the additive nonignorable model can be viewed as an alternative to imposing one of those two missingness mechanisms, and instead letting the data determine a compromise.

The existence results of Hirano et al. (1998, 2001) are limited to distributions that are absolutely continuous with respect to the product Lebesgue measure (Hirano et al., 2001) or that have finite support (Hirano et al., 1998), meaning that they do not cover problems with variables of mixed type. Our goal in this section is to extend their identification results to general variable types, and to permit the usage of auxiliary marginal information that can be coarser compared to $f(y)$, as in Definition 5, in which case β in (2) will be restricted to be in $\langle \mathcal{U} \rangle$.

To clearly separate identification from estimation issues, in our identification results we assume the observed-data density $f(x, y^*)$ to be known and availability of perfect auxiliary marginal information, corresponding to knowing the value of $E[u(Y)]$ for all functions $u \in \mathcal{U}$ as in Definition 5. Using a pattern mixture model formulation for missing data, the observed-data density $f(x, y^*)$ in this case corresponds to $\pi f(x \mid M = 1)$ when $M = 1$, and $(1 - \pi) f(x, y \mid M = 0)$ when $M = 0$, where π is the probability of $M = 1$. For each $u \in \mathcal{U}$ we also can derive each $E[u(Y) \mid M = 0] = \int u(y) f(y \mid M = 0) \mu(dy)$ using $f(y \mid M = 0) = \int f(x, y \mid M = 0) \nu(dx)$ from the observed-data distribution. Combining these with auxiliary marginal information, we can obtain $E[u(Y) \mid M = 1] = \int u(y) f(y \mid M = 1) \mu(dy) = \{E[u(Y)] - (1-\pi) E[u(Y) \mid M = 0]\}/\pi$. This means that auxiliary marginal information allows us to find the value of integrals which are computed with respect to the distribution of $Y \mid M = 1$ that cannot be obtained from the observed-data distribution. Therefore, while $f(x, y \mid M = 1)$ is unknown, we know its marginal $f(x \mid M = 1)$ and have a set of constraints given by the values of the integrals $\int u(y) f(y \mid M = 1) \mu(dy)$ for $u \in \mathcal{U}$.

From an information theoretic point of view, it is natural to think of approximating the true $f(x, y \mid M = 1)$ by an information projection of $f(x, y \mid M = 0)$ onto the set of distributions that have the X-marginal given by $f(x \mid M = 1)$ and that also satisfy the constraints imposed by the auxiliary marginal information. The $\mathtt{f}$-divergence $I_{\mathtt{f}}$ between distributions with densities $g^*(x, y)$ and $g(x, y)$ is given by

$$I_{\mathtt{f}}(g^*, g) = \iint \mathtt{f}\left[\frac{g^*(x, y)}{g(x, y)}\right] g(x, y) \nu(dx) \mu(dy),$$

for a convex and differentiable function $\mathfrak{f} : (0, \infty) \mapsto \mathbb{R}$ (Csiszár, 1963). For example, when $\mathfrak{f}(z) = z\log(z)$ we obtain the Kullback–Leibler divergence. The $\mathfrak{f}$-projection of a probability distribution with density $g(x, y)$ onto a set of probability distributions $\mathcal{C}$ is defined as the element in $\mathcal{C}$ with density $g^*(x, y)$ that minimizes $I_{\mathfrak{f}}(g^*, g)$ (see, e.g., Liese and Vajda, 1987, Ch. 8). We will show that there is an intrinsic connection between the function $\mathfrak{f}$ used in $\mathfrak{f}$-projections and the link function λ used to define additive nonignorable missingness mechanisms. Our results rely on those of Liese and Vajda (1987), which require $\mathfrak{f}$ not to increase too fast as its argument approaches infinity.

Assumption 1 (Regular link function). *The link function $\lambda : (0, 1) \mapsto \mathbb{R}$ is differentiable and monotonically increasing.*

Assumption 2 (Growth of $\mathfrak{f}$). *The function $\mathfrak{f}$ satisfies the property that for every $t > 1$ there exist positive constants t_0, t_1, t_2, t_3 such that for all $z > t_0$, $\mathfrak{f}(tz) \leq t_1\mathfrak{f}(z) + t_2 z + t_3$.*

Theorem 1 (Identification). *Let X be a vector of always observed random variables, Y be a random variable subject to missingness and M be its missingness indicator. Assume that the observed-data density $f(x, y^*)$ and auxiliary marginal information $\{E[u(Y)]\}_{u\in\mathcal{U}}$ are derived from a distribution that satisfies $\lambda[f(M = 1 \mid x, y)] = \alpha(x) + \beta(y)$, where λ satisfies Assumption 1, $\alpha \in L_1\{f(x \mid M = 1)\}$, and $\beta \in \langle\mathcal{U}\rangle$. Assume that $\mathfrak{f}_\lambda(z) = \int_0^z \lambda[v/(c+v)]dv$, $c = (1 - \pi)/\pi$, satisfies Assumption 2. Then $f(x, y \mid M = 1)$ is the $\mathfrak{f}_\lambda$-projection of $f(x, y \mid M = 0)$ onto the set of distributions that match both the marginal defined by $f(x \mid M = 1)$ and the expectations given by $\{E[u(Y) \mid M = 1]\}_{u\in\mathcal{U}}$.*

All of our proofs are presented in Appendix 2. They rely on some results on $\mathfrak{f}$-projections presented in Appendix 1. Theorem 1 indicates that under additive nonignorability, if the missingness mechanism satisfies $\lambda[f(M = 1 \mid x, y)] = \alpha(x) + \beta(y)$, with $\alpha \in L_1\{f(x \mid M = 1)\}$ and $\beta \in \langle\mathcal{U}\rangle$, then the full-data distribution can be obtained from its implied observed-data distribution and from the auxiliary marginal information represented by $\{E[u(Y)]\}_{u\in\mathcal{U}}$, since we only need these pieces to recover $f(x, y \mid M = 1)$ and thereby $f(x, y, m)$. We are of course assuming that α and β are adequately set-up to ensure full-data identifiability of the class containing $f(x, y, m)$, as mentioned in Section 2.2. Theorem 1 also requires $\mathfrak{f}_\lambda$ to satisfy Assumption 2, for which a sufficient condition is that $\lim_{z\to\infty} \mathfrak{f}(z)/z^a = 0$ for some

$a > 0$ (Liese and Vajda, 1987, p. 171). This is the case for common link functions, such as the logit, probit, complementary log-log, and the link functions proposed by Aranda-Ordaz (1981), just to name some, for all of which $\lim_{z\to\infty} \mathfrak{f}_\lambda(z)/z^2 = 0$.

Theorem 1 indicates the largest additive nonignorable missingness mechanism that we can identify given the available auxiliary marginal information. If $E(Y)$ is all we have access to, this result says that the model $\lambda[f(M = 1 \mid x, y)] = \alpha(x) + by$, with $b \in \mathbb{R}$, is identifiable; that is, the missingness mechanism can include a main linear effect of Y, such as in the example 1.9 of Little and Rubin (2002). If $\mathcal{U} = \{u_1, \ldots, u_k\}$, then the model $\lambda[f(M = 1 \mid x, y)] = \alpha(x) + \sum_{j=1}^{k} b_j u_j(y)$ is identifiable. If we know the marginal distribution of Y, then $\mathcal{U}$ can be taken as the set of all integrable functions, and this result says that the model $\lambda[f(M = 1 \mid x, y)] = \alpha(x) + \beta(y)$, with $\beta \in L_1\{f(y \mid M = 1)\}$, is identifiable, which corresponds to the additive nonignorable mechanism of Hirano et al. (1998, 2001).

The statement of Theorem 1 suggests a plug-in approach for obtaining an estimate of $f(x, y \mid M = 1)$, by computing the $\mathfrak{f}$-projection of an estimate of $f(x, y \mid M = 0)$ onto the set of distributions that match estimates of $f(x \mid M = 1)$ and $\{E[u(Y) \mid M = 1]\}_{u\in\mathcal{U}}$. While there exist algorithms for doing so (e.g., Rüschendorf, 1995; Bhattacharya, 2006), they can be challenging to implement as they require iterative approximation of potentially complex integrals (Rüschendorf, 1995). Instead, we rely on the theory of $\mathfrak{f}$-projections merely for identification results, and present a more straightforward, likelihood-based implementation in Section 4.

The following result indicates that a full-data distribution derived assuming additive nonignorability is observationally equivalent to the true full-data distribution.

Theorem 2 (Nonparametric identification). *Let the observed-data density $h(x, y^*)$ and the auxiliary marginal information $\{\int u(y)h(y)\mu(dy)\}_{u\in\mathcal{U}}$ be derived from a full-data density $h(x, y, m)$. Let $\mathfrak{f}_\lambda(z) = \int_0^z \lambda[v/(c + v)]dv$, with $c = (1 - \pi)/\pi$ and $\pi = h(M = 1)$, satisfy Assumption 2, for a function λ satisfying Assumption 1. Let $g(x, y \mid M = 1)$ denote the $\mathfrak{f}_\lambda$-projection of $h(x, y \mid M = 0)$ onto the set of distributions that match the marginal*

defined by $h(x \mid M = 1)$ *and the integrals* $\{\int u(y)h(y \mid M = 1)\mu(dy)\}_{u \in \mathcal{U}}$. *Define a full-data distribution as* $g(x, y, m) = \{g(x, y \mid M = 1)\pi\}^m \{h(x, y, M = 0)\}^{1-m}$. *Then* $g(x, y, m)$ *encodes an additive nonignorable missingness mechanism with* $\lambda[g(M = 1 \mid x, y)] \in \overline{\langle L_1\{h(x \mid M = 1)\} \cup \mathcal{U}\rangle}$, *and furthermore* $g(x, y^*) = h(x, y^*)$ *and* $\int u(y)g(y)\mu(dy) = \int u(y)h(y)\mu(dy)$ *for all* $u \in \mathcal{U}$.

This result indicates that if one derives a full-data density $g(x, y, m)$ assuming additive nonignorability from a given observed-data density $h(x, y^*)$ and auxiliary marginal information $\{\int u(y)h(y)\mu(dy)\}_{u \in \mathcal{U}}$, then $g(x, y, m)$ implies back the original $h(x, y^*)$ and $\{\int u(y)h(y)\mu(dy)\}_{u \in \mathcal{U}}$. In other words, additive nonignorability induces a one-to-one mapping from the set of observed-data distributions and auxiliary marginal information to the set of full-data distributions. A technical detail in Theorem 2 is that $\lambda[g(M = 1 \mid x, y)]$ need not be in $\langle L_1\{h(x \mid M = 1)\} \cup \mathcal{U}\rangle$ but it could be a limit point outside of this set, although in that case $\lambda[g(M = 1 \mid x, y)]$ can be arbitrarily approximated by functions of the form $\alpha(x) + \beta(y)$ where $\alpha \in L_1\{h(x \mid M = 1)\}$, and $\beta \in \langle \mathcal{U}\rangle$.

3.2 Multivariate nonresponse

We now extend the concept of additive nonignorability to the context of multivariate item nonresponse, where each of the variables in $Y = (Y_1, \ldots, Y_p)$ is subject to nonresponse, with $M = (M_1, \ldots, M_p)$ being its vector of missingness indicators. For now, we still consider the vector of variables X to be fully observed, but we relax this requirement in Section 4.2.

We begin by defining a comprehensive class of sequentially additive nonignorable missingness mechanisms that allows M_j to depend directly on Y_j for each j, yet also meets the criterion of nonparametric identification. In some contexts, however, analysts may find it convenient to use submodels of the comprehensive version that we introduce. These may be easier to interpret or estimate, as we discuss in Section 3.3 and Section 4. The identification results for the comprehensive version, however, provide assurance that its submodels also are identifiable, albeit without the advantages endowed by nonparametric identification.

We factorize the missingness mechanism as

$$f(m_1, \ldots, m_p \mid x, y) = \prod_{j=1}^{p} f(m_j \mid x, y, m_{<j}),$$

where we let $f(m_j \mid x, y, m_{<j})$ be as general as possible to obtain nonparametric identification. A similar sequential factorization strategy is used by Ibrahim et al. (1999), among others. It requires us to impose an ordering on the p variables. To facilitate explanations, we proceed as if the variables $Y_1, \ldots, Y_p$ are indexed by the order in which they are collected, which is a natural choice in longitudinal studies or when we know the order in which questions are administered in a survey. Of course, sequential additive nonignorability can be defined for any other ordering, and different orderings will lead to different missingness mechanisms. We discuss guidelines for selecting orderings in Section 4.1. The order of the X variables with respect to those in Y is irrelevant.

To motivate the comprehensive version of the sequentially additive nonignorable missingness mechanism, we shall think of a hypothetical respondent from whom we attempt to collect values of the variables $Y_1, \ldots, Y_p$. We start by trying to collect her value of Y_1, but she may or may not report it. Whether she reports it or not is determined by a probabilistic mechanism $f(m_1 \mid x, y)$, which we assume to satisfy

$$\lambda\{f(M_1 = 1 \mid x, y)\} = \alpha_1(x, y_{>1}) + \beta_1(y),$$

for some functions α_1 and β_1 subject to constraints described later. This indicates that, given a value of $Y_{>1}$, the nonresponse for Y_1 follows an additive nonignorable mechanism as in (2). The functions $\alpha_1(x, y_{>1})$ and $\beta_1(y) \equiv \beta_1(y_1, y_{>1})$ represent interactions between X and $Y_{>1}$, and Y_1 and $Y_{>1}$, respectively, but the model does not allow interactions between Y_1 and X. In particular, this means that the nonresponse for Y_1 can depend on Y_1, and this dependence can change across the values of $Y_{>1}$ but is homogeneous across the values of X. The result of our attempt to measure Y_1 is a realization of its materialized variable Y_1^*.

We then attempt to measure the respondent's value of Y_2. Whether she reports this value is determined by a probabilistic mechanism that we assume to satisfy $f(m_2 \mid x, y, m_1) =$

$f(m_2 \mid x, y_1^*, y_{\geq 2})$, that is, the probability of nonresponse for Y_2 depends on Y_1 and M_1 only through the materialized variable Y_1^*, namely, if the value of Y_1 is not revealed then it does not influence the probability of nonresponse for Y_2. Since Y_1^* captures all the information of M_1, the probability of nonresponse for Y_2 does depend on whether Y_1 is reported. We further assume that

$$\lambda\{f(M_2 = 1 \mid x, y_1^*, y_{\geq 2})\} = \alpha_2(x, y_1^*, y_{>2}) + \beta_2(y_{\geq 2}),$$

for some functions α_2 and β_2 described later. Similarly as for the first item, for each value of $Y_{>2}$ the nonresponse for Y_2 follows an additive nonignorable mechanism, where $\alpha_2(x, y_1^*, y_{>2})$ represents interactions between (X, Y_1^*) and $Y_{>2}$, or equivalently, interactions between (X, Y_1, M_1) and $Y_{>2}$ which are homogeneous across the missing values of Y_1. The direct dependence of M_2 on Y_2 is captured by $\beta_2(y_{\geq 2}) \equiv \beta_2(y_2, y_{>2})$, which allows this dependence to vary with $Y_{>2}$. The dependence of M_2 on Y_2, however, is homogeneous across the values of (X, Y_1, M_1). Thus far, the result of our data collection process is (Y_1^*, Y_2^*).

After having attempted to collect the respondent's values for the first $j-1$ variables, $Y_{<j}$, we have actually obtained a realization of their materialized variables $Y_{<j}^*$. At this point, the missingness mechanism for whether we observe the respondent's value of Y_j is defined by $f(m_j \mid x, y, m_{<j}) = f(m_j \mid x, y_{<j}^*, y_{\geq j})$. The assumption in this mechanism is that the nonresponse for Y_j does not depend on the missing values of the previous variables in the sequence, that is, its dependence on $Y_{<j}$ and $M_{<j}$ comes only through the materialized variables $Y_{<j}^*$. We further assume that, for each value of $Y_{>j}$ the nonresponse mechanism for Y_j is additive nonignorable, that is,

$$\lambda\{f(M_j = 1 \mid x, y_{<j}^*, y_{\geq j})\} = \alpha_j(x, y_{<j}^*, y_{>j}) + \beta_j(y_{\geq j}).$$

Here, the function $\alpha_j(x, y_{<j}^*, y_{>j})$ represents interactions between $(X, Y_{<j}, M_{<j})$ and $Y_{>j}$, although these interactions are constant across the missing values of $Y_{<j}$. Also, the function $\beta_j(y_{\geq j}) \equiv \beta_j(y_j, y_{>j})$ represents interactions between Y_j and $Y_{>j}$, but the model does not have an interaction between Y_j and $(X, Y_{<j}, M_{<j})$.

The final step in our data collection attempt is to try to record the respondent's value of Y_p, at which point we have collected her value of $Y^*_{<p}$. For the final variable, we assume the missingness mechanism to be $f(m_p \mid x, y, m_{<p}) = f(m_p \mid x, y^*_{<p}, y_p)$, with

$$\lambda\{f(M_p = 1 \mid x, y^*_{<p}, y_p)\} = \alpha_p(x, y^*_{<p}) + \beta_p(y_p),$$

meaning that the nonresponse for the last variable does not depend on any of the missing values for the previous variables, but it can depend on the value of Y_p itself, although this dependence is homogeneous across all the values of $(X, Y_{<p}, M_{<p})$.

Definition 6 (Sequential additive nonignorability). *Let X be a vector of always observed random variables, Y be a vector of p random variables subject to missingness and M be its vector of missingness indicators. A missingness mechanism is sequentially additive nonignorable if it can be written as $f(m \mid x, y) = \prod_{j=1}^{p} f(m_j \mid x, y, m_{<j})$, where for $j = 1, \ldots, p$,*

$$f(m_j \mid x, y, m_{<j}) = f(m_j \mid x, y^*_{<j}, y_{\geq j}), \tag{3}$$

with

$$\lambda\{f(M_j = 1 \mid x, y^*_{<j}, y_{\geq j})\} = \alpha_j(x, y^*_{<j}, y_{>j}) + \beta_j(y_{\geq j}), \tag{4}$$

where λ is a link function, and α_j and β_j are real-valued functions.

The α_j and β_j functions in this definition require some constraints to guarantee full-data identifiability. In this article we constrain $\beta_j(y^0_j, y_{>j}) = 0$ for some arbitrary value y^0_j of Y_j and for all values $y_{>j}$ of $Y_{>j}$, while leaving the α_j functions unconstrained; this restriction implies that $\beta_j(y_{\geq j})$ cannot be expressed with additive terms that only depend on $y_{>j}$. Other restrictions are possible; for example, one could add an intercept to the linear part in (4) while also constraining α_j. In Theorems 3 and 4 we impose further restrictions on the α_j and β_j functions to guarantee identifiability and nonparametric identifiability. The restrictions for β_j are determined by the auxiliary marginal information on the distribution of $Y_{\geq j}$, specifically we require $\beta_j \in \langle \mathcal{U}_{\geq j} \rangle$, where $\mathcal{U}_{\geq j}$ denotes the set of functions in $\mathcal{U}$ that depend exclusively on $y_{\geq j}$.

Our proofs of identifiability and nonparametric identifiability rely on Algorithm 1, which uses a sequence of information projections to construct a full-data distribution that satisfies sequential additive nonignorability, taking the observed-data density $f(x, y^*)$ and the auxiliary marginal information $\{E[u(Y)]\}_{u\in\mathcal{U}}$ as input. The true full-data density, from which $f(x, y^*)$ and $\{E[u(Y)]\}_{u\in\mathcal{U}}$ are derived, is denoted $f(x, y, m)$, and the densities obtained from Algorithm 1 are denoted with g. Theorem 3 shows identifiability, since if $f(x, y, m)$ truly satisfies sequential additive nonignorability, then the output $g(x, y, m)$ of Algorithm 1 equals $f(x, y, m)$ almost surely. Theorem 4 shows nonparametric identifiability, since sequential additive nonignorability cannot be refuted using $f(x, y^*)$ and $\{E[u(Y)]\}_{u\in\mathcal{U}}$ alone, given that $g(x, y, m)$ is observationally equivalent to the true $f(x, y, m)$.

Algorithm 1. *Full-data distribution construction algorithm.*

Input $g(x, y^*) \equiv f(x, y^*)$, $\{E[u(Y)]\}_{u\in\mathcal{U}}$.

For $j = p, \ldots, 1$

a. *Use* $g(x, y^*_{\leq j}, y_{>j})$ *to derive* $g(x, y^*_{<j}, y_{\geq j} \mid M_j = 0)$, $g(x, y^*_{<j}, y_{>j} \mid M_j = 1)$, *and* $g(y_{\geq j}, M_j = 0)$, *and* $\pi_j = g(M_j = 1)$.

b. *For each* $u \in \mathcal{U}_{\geq j}$ *compute*

$$E_g[u(Y_{\geq j}) \mid M_j = 1] = \{E[u(Y_{\geq j})] - \int u(y_{\geq j}) g(y_{\geq j}, M_j = 0)\mu(dy_{\geq j})\}/\pi_j$$

c. *Find* $g(x, y^*_{<j}, y_{\geq j} \mid M_j = 1)$ *as the* $\mathbb{f}_{\lambda,j}$*-projection of* $g(x, y^*_{<j}, y_{\geq j} \mid M_j = 0)$ *onto the set of distributions that match the marginal* $g(x, y^*_{<j}, y_{>j} \mid M_j = 1)$ *and the expectations* $\{E_g[u(Y_{\geq j}) \mid M_j = 1]; u \in \mathcal{U}_{\geq j}\}$,
with $\mathbb{f}_{\lambda,j}(z) = \int_0^z \lambda[v/(c_j + v)]dv$, $c_j = (1 - \pi_j)/\pi_j$.

d. *Obtain* $g(x, y^*_{<j}, y_{\geq j}) = \sum_{m_j=0}^{1} g(x, y^*_{<j}, y_{\geq j} \mid M_j = m_j)\pi_j^{m_j}(1 - \pi_j)^{1-m_j}$, *and* $g(m_j \mid x, y^*_{<j}, y_{\geq j}) = \frac{g(x, y^*_{<j}, y_{\geq j} \mid M_j = m_j)}{g(x, y^*_{<j}, y_{\geq j})}\pi_j^{m_j}(1 - \pi_j)^{1-m_j}$.

Output $g(x, y, m) = g(x, y)\prod_{j=1}^{p} g(m_j \mid x, y^*_{<j}, y_{\geq j})$.

Theorem 3 (Identification). *Let* X *be a vector of always observed random variables,* Y *be a random vector subject to missingness and* M *be its vector of missingness indicators. Assume that the observed-data density* $f(x, y^*)$ *and auxiliary marginal information* $\{E[u(Y)]\}_{u\in\mathcal{U}}$ *are derived from a distribution with density* $f(x, y, m)$ *that encodes a sequentially additive nonignorable missingness mechanism as in Definition 6, where* λ *satisfies*

Assumption 1, $\alpha_j \in L_1\{f(x, y^*_{<j}, y_{>j} \mid M_j = 1)\}$, and $\beta_j \in \langle \mathcal{U}_{\geq j} \rangle$ for all $j = 1, \ldots, p$. Assume that each $\mathtt{f}_{\lambda,j}(z) = \int_0^z \lambda[v/(c_j + v)]dv$, with $c_j = (1 - \pi_j)/\pi_j$ and $\pi_j = f(M_j = 1)$, satisfies Assumption 2. Then

1. *$f(x, y^*_{<j}, y_{\geq j} \mid M_j = 1)$ is the $\mathtt{f}_{\lambda,j}$-projection of $f(x, y^*_{<j}, y_{\geq j} \mid M_j = 0)$ onto the set of distributions that match the marginal defined by $f(x, y^*_{<j}, y_{>j} \mid M_j = 1)$ and the expectations given by $\{E[u(Y_{\geq j}) \mid M_j = 1]; u \in \mathcal{U}_{\geq j}\}$, for all $j = 1, \ldots, p$.*

2. *The output $g(x, y, m)$ of Algorithm 1 equals $f(x, y, m)$ almost surely.*

Theorem 4 (Nonparametric identification). *Let the observed-data density $h(x, y^*)$ and the auxiliary marginal information $\{\int u(y)h(y)\mu(dy)\}_{u \in \mathcal{U}}$ be derived from a full-data density $h(x, y, m)$. Let λ satisfy Assumption 1. Let $\mathtt{f}_{\lambda,j}(z) = \int_0^z \lambda[v/(c_j + v)]dv$, with $c_j = (1 - \pi_j)/\pi_j$ and $\pi_j = h(M_j = 1)$, satisfy Assumption 2 for each $j = 1, \ldots, p$. Let $g(x, y, m)$ be constructed as in Algorithm 1. Then*

1. *$g(x, y, m)$ encodes a sequentially additive nonignorable missingness mechanism with $\lambda[g(M_j = 1 \mid x, y^*_{<j}, y_{\geq j})] \in \overline{\langle L_1\{g(x, y^*_{<j}, y_{>j} \mid M_j = 1)\} \cup \mathcal{U}_{\geq j} \rangle}$, $j = 1, \ldots, p$.*

2. *The output $g(x, y, m)$ of Algorithm 1 and $h(x, y, m)$ are observationally equivalent, that is $g(x, y^*) = h(x, y^*)$ almost surely, and $\int u(y)g(y)\mu(dy) = \int u(y)h(y)\mu(dy)$ for all $u \in \mathcal{U}$.*

Algorithm 1 suggests a plug-in implementation, starting from estimates of the observed-data distribution and the auxiliary marginal information. However, as we previously mentioned, we only use the theory of information projections for our identification results, and provide likelihood-based implementations in Section 4. Therefore, we use Algorithm 1 only as a theoretical tool for guaranteeing identification and nonparametric identification under sequential additive nonignorability.

3.3 Applying the identification results

In some settings, it may be difficult to conceptualize a response process that fully corresponds to the comprehensive version of sequential additive nonignorability. However, as we illustrate in this section, the comprehensive version includes important special cases of missingness mechanisms that are readily amenable to interpretation. The identification results in Section 3 assure analysts that models using these special-case missingness mechanisms can be estimated with enough data plus auxiliary marginal information. In contrast, models that violate the constraints in (3) or (4) may not be identifiable even with infinite amounts of data. Thus, analysts can use the identification results to specify missingness mechanisms that are interpretable submodels and know that these are identifiable. Furthermore, analysts can work with the comprehensive version or with a large subclass of sequentially additive nonignorable models to enable data-driven compromises between simpler cases.

As with the univariate additive nonignorable mechanism, the multivariate mechanism encodes ignorable and nonignorable cases. When we set $\beta_j(y_{\geq j}) = 0$ and $\alpha_j(x, y^*_{<j}, y_{>j}) = \alpha_j(x, y^*_{<j})$ for all j, we have an ignorable missingness mechanism as it does not depend on missing values. When we set $\alpha_j(x, y^*_{<j}, y_{\geq j}) = \alpha_j$ and $\beta_j(y_{\geq j}) = \beta_j(y_j)$ for all j, that is, nonresponse for Y_j depends only on the value of Y_j itself, we encode a multivariate version of the selection model of Hausman and Wise (1979). Combining these two cases, taking $\alpha_j(x, y^*_{<j}, y_{>j}) = \alpha_j(x, y^*_{<j})$ and $\beta_j(y_{\geq j}) = \beta_j(y_j)$ for all j, corresponds to the attrition mechanism of Hoonhout and Ridder (2018) in the case of monotone nonresponse. This also encodes an identifiable nonignorable nonresponse process with nonmonotone nonresponse. An example where this special case of the comprehensive mechanism could be plausible is when $Y_1, \ldots, Y_p$ are a set of medical tests that are administered sequentially. Here, $M_j = 0$ indicates that test j was performed on the patient, and $M_j = 1$ indicates otherwise. For some tests, we may have marginal distributions of the outcomes, say from baseline studies or meta-analyses. It is reasonable to assume that physicians might not administer test j on some patient because she is extremely confident that the patient's Y_j would be in a medically safe range. However, the physician's decision to administer the test also could depend on

demographic characteristics X and on values of previously administered tests $Y^*_{<j}$. It seems plausible that the potential results of tests that were not previously administered have no influence on the physician's decision to administer a new test.

A further simplification of the previous subclass can be obtained by seting $\alpha_j(x, y^*_{<j}, y_{>j}) = \alpha_j(x)$ and $\beta_j(y_{\geq j}) = \beta_j(y_j)$ for all j. This allows nonresponse for Y_j to depend on its values, as well as the variables in X, while ensuring that inferences are invariant to the ordering of $Y_1, \ldots, Y_p$. This can be convenient when analysts have no reasonable assumptions on which to base an ordering. Fully observed variables in Y can be put in any order with respect to those with missingness, given that we do not model their missingness mechanism. When some of the Y variables are fully observed, say without loss of generality $Y_{\geq j'}$, the submodel where $\alpha_j(x, y^*_{<j}, y_{>j}) = \alpha_j(x, y_{\geq j'})$ and $\beta_j(y_{\geq j}) = \beta_j(y_j, y_{\geq j'})$, for $j < j'$, is also invariant to the ordering of the variables. An illustration of this situation appears in the example of Section 5.

Another feature of the general missingness mechanism is that nonresponse for a variable Y_j can, although it does not have to, depend on the following variables in the sequence $Y_{>j}$. For example, suppose a survey collects, among other items, Y_j which indicates whether voted in the last election, and $Y_{j'}$ which indicates political affiliation; here, $j < j'$. Marginal information about these variables is available from external sources. It seems plausible that the probability of reporting voting turnout might depend on the political affiliation and on whether the respondent voted, regardless of whether these variables get reported. An advantage of sequentially additive nonignorable mechanisms is that such associations can be picked up if they exist.

Finally, an interesting connection is obtained by reversing the order of the items $Y_1, \ldots, Y_p$ and fixing $\beta_j(y_{\geq j}) = 0$ for all j, leading to the permutation missingness mechanism of Robins (1997). That mechanism says that given an ordering of the study variables, the nonresponse propensity for variable j depends on the values of the previous study variables in the order, whether observed or not, but not on variable j nor on the following missing values in the order, which is the reverse of the interpretation that we have given. Robins

(1997) emphasized, however, that a limitation of the permutation missingness mechanism is that it does not allow the probability of missingness for a particular variable to depend on the value of that variable. With sequential additive nonignorability, we are not subject to this limitation as the auxiliary marginal information allows us to obtain $\beta_j(y_{\geq j}) \neq 0$.

As we can see, sequential additive nonignorability leads to very flexible classes of missingness models, as it encompasses a number of important particular cases that encode potentially plausible missingness mechanisms.

4 Implementation

4.1 Practical considerations

The particular ordering of the variables encodes assumptions about the missingness mechanism and hence affects inferences. Of course, orderings imply distributional assumptions for any multivariate missing data modeling strategy based on chained conditional distributions (Ibrahim et al., 1999; Xu et al., 2016). Depending on the context, some orderings may lead to assumptions deemed more plausible than others. In some contexts, it may be reasonable to order variables temporally, for example, following the sequence in which questions are asked in a survey questionnaire or time points in a longitudinal study. In other settings, natural orderings may not be apparent. In this case, analysts can view the comprehensive version of sequential additive nonignorability as a rich class of multivariate nonignorable missing data models that allow M_j to depend on Y_j, as well as allow estimation of additional dependencies that the analyst may not have considered initially. In such cases, it may be computationally convenient to order variables from highest to lowest fractions of missing data, so that the richest models are used for the variables with the most missing data. Alternatively, analysts could use the simpler mechanism described in Section 3.3 that does not require any ordering. When the ordering is somewhat arbitrary, it is prudent for analysts to analyze the sensitivity of results to different orderings, as we do in Section 5.

Furthermore, in practice one works with finite samples, and therefore working with models with the most comprehensive structure in (4) may lead to complications. For example, if some category of a categorical variable, or a combination of categories of different categorical variables, happens not to be observed in the sample, then their corresponding parameters are not estimable from the likelihood function alone, as it will be constant as a function of those parameters. In such cases, maximum likelihood estimates will not be unique, and Bayesian inference reliant on Markov chain Monte Carlo will suffer from convergence issues, unless strongly informative priors are imposed on those parameters. Similar issues would occur if the number of unique model parameters exceeds the sample size. Thus, in many practical circumstances it is prudent to work with a model that respects the form of (4) but that does not include all possible interactions within the variables in $(X, Y^*_{<j}, Y_{>j})$ nor within the variables in $Y_{\geq j}$.

4.2 Partially ignorable multivariate nonresponse

We now return to the general set-up introduced in Section 2, where some or all of the variables in the vector X, for which we do not have auxiliary marginal information, may also be subject to missingness, with W denoting its vector of missingness indicators. In such cases, assuming sequential additive nonignorability for the missingness in both Y and X would lead to nonidentifiable models, since we only have auxiliary marginal information for Y. Therefore, to implement sequential additive nonignorability in those situations, we assume partial ignorability of the missingness mechanism (Harel and Schafer, 2009). Specifically, we can write the missingness mechanism as

$$f(w, m \mid x, y) = f(m \mid x, y) f(w \mid x, y, m), \tag{5}$$

where we assume $f(m \mid x, y)$ to be sequentially additive nonignorable, and we assume the missingness of the X variables to be partially missing always at random (Harel and Schafer, 2009; Mealli and Rubin, 2015), that is,

$$f(w \mid x, y, m) = f(w \mid x_{\bar{w}}, y_{\bar{m}}, m) \equiv f(w \mid x_{\bar{w}}, y^*), \tag{6}$$

for all $w \in \{0,1\}^q$ and $m \in \{0,1\}^p$. This assumption indicates that the probability of observing missingness pattern w in the X variables does not depend on the unobserved values in X and Y. We take (6) as an assumption on the missingness mechanism and not only on a specific realized value w. For related discussions see Seaman et al. (2013) and Mealli and Rubin (2015).

The assumption in (6) can be decomposed into two parts. First, $f(w \mid x, y, m) = f(w \mid x, y^*)$, which says that this missingness mechanism is homogeneous across the missing values of the Y variables. In such case, Sadinle and Reiter (2018) showed that if the missingness mechanism $f(w \mid x, y^*)$ leads to nonparametric identified $f(x, w \mid y^*)$ for each y^*, and if $f(m \mid x, y)$ leads to nonparametric identified $f(x, y, m)$, then the combined missingness mechanism $f(m \mid x, y)f(w \mid x, y^*)$ leads to a nonparametric identified full-data distribution $f(x, y, w, m)$. The second part of the assumption in (6) is $f(w \mid x, y^*) = f(w \mid x_{\bar{w}}, y^*)$, which is a missing always at random assumption conditional on Y^*. Gill et al. (1997) showed that the missing always at random assumption leads to nonparametric identified distributions. Therefore, following Sadinle and Reiter (2018), we obtain the property of nonparametric identification for full-data distributions derived under (5) and (6), with $f(m \mid x, y)$ being sequentially additive nonignorable.

4.3 Likelihood-based inference

We consider the scenario where our initial goal is to use a random sample $\{(x_i, y_i)\}_{i=1}^n$ from a distribution with density $f(x \mid y, \theta)f(y \mid \kappa)$ to draw inferences on parameter vectors θ and κ. Here, the auxiliary marginal information about the distribution of Y is included via the parameters κ. If the sample is subject to missingness, we instead think of a full-data random sample $\{(x_i, y_i, w_i, m_i)\}_{i=1}^n$ drawn from a full-data distribution with density

$$f(w \mid x, y, m, \phi)f(m \mid x, y, \gamma)f(x \mid y, \theta)f(y \mid \kappa) \equiv \ell(\phi, \gamma, \theta, \kappa;\ x, y, w, m), \tag{7}$$

with ϕ and γ parameterizing the missingness mechanism. The full-data likelihood function is therefore $L(\phi, \gamma, \theta, \kappa) = \prod_{i=1}^n \ell(\phi, \gamma, \theta, \kappa;\ x_i, y_i, w_i, m_i)$. The full-data random sample gets

materialized as an observed-data random sample $\{(x_{i,\bar{w}_i}, y_{i,\bar{m}_i}, w_i, m_i)\}_{i=1}^n \equiv \{(x_i^*, y_i^*)\}_{i=1}^n$. The observed-data likelihood function is derived by integrating L over the missing values y_{i,m_i} and x_{i,w_i}, according to each missingness pattern m_i and w_i, that is, $L_{obs}(\phi, \gamma, \theta, \kappa) = \prod_{i=1}^n \ell_{obs}(\phi, \gamma, \theta, \kappa;\ x_{i,\bar{w}_i}, y_{i,\bar{m}_i}, w_i, m_i)$, with

$$\ell_{obs}(\phi, \gamma, \theta, \kappa;\ x_{\bar{w}}, y_{\bar{m}}, w, m) = \iint \ell(\phi, \gamma, \theta, \kappa;\ x, y, w, m)\mu(dy_m)\nu(dx_w).$$

It is easy to check that taking $f(w \mid x, y, m, \phi)$ in (7) to be partially missing always at random, as in (6), implies that this part of the missingness mechanism can be ignored from the likelihood function for inference on γ, θ and κ. We therefore work with the likelihood function $L_{obs}(\gamma, \theta, \kappa) = \prod_{i=1}^n \ell_{obs}(\gamma, \theta, \kappa;\ x_{i,\bar{w}_i}, y_{i,\bar{m}_i}, w_i, m_i)$, where

$$\ell_{obs}(\gamma, \theta, \kappa;\ x_{\bar{w}}, y_{\bar{m}}, w, m) = \iint f(m \mid x, y, \gamma) f(x \mid y, \theta) f(y \mid \kappa)\mu(dy_m)\nu(dx_w).$$

Taking $f(m \mid x, y, \gamma)$ as sequentially additive nonignorable permits us to write it as a product of logistic regressions $\prod_{j=1}^p f(m_j \mid x, y, m_{<j}, \gamma_j)$, where

$$\lambda[f(M_j = 1 \mid x, y, m_{<j}, \gamma_j)] = \alpha_j(x, y_{<j}^*, y_{>j}) + \beta_j(y_{\geq j}), \tag{8}$$

with γ_j representing the parameter functions α_j and β_j. The nature of these functions depends on the type of variables in X and Y, and on the auxiliary marginal information on the distribution of Y. If we know $f(y)$, in the case of having only categorical variables in X and Y, under the constraints $\beta_j(y_j^0, y_{>j}) = 0$ for arbitrary y_j^0 and for all $y_{>j}$, we have that the value of α_j for each possible value of $(X, Y_{<j}^*, Y_{>j})$ and the value of β_j for each possible value of $(Y_j, Y_{>j})$, $Y_j \neq y_j^0$, represent the parameters of the full model (8). While this model is nonparametrically identified, in practice we may encounter issues estimating its parameters due to the finiteness of the sample, as we discuss in Section 4.1. If X and Y contain continuous variables, the α_j and β_j functions could be modeled using splines or Gaussian processes (e.g. Choudhuri et al., 2007), which although not strictly nonparametric can be flexible enough to capture complex distributional features. If the auxiliary marginal information is simply a finite set of moment restrictions $\{E[u(Y)]\}_{u \in \mathcal{U}}$ for $\mathcal{U} = \{u_1, \ldots, u_k\}$, then it is easy to specify $\beta(y) = \sum_{j=1}^k b_j u_j(y)$.

Our final working likelihood is $L_{obs}(\gamma, \theta, \kappa)A(\kappa)$, where $A(\kappa)$ is a function whose form depends on the nature of the auxiliary marginal information. If we have access to the true κ, for example if Y is categorical and we know its true distribution from a census, then $A(\kappa)$ is simply an indicator function that equals zero when κ is different from its census value. If we have access to an additional fully observed random sample $\{y_i\}_{i=n+1}^{m}$ from the distribution of Y, as with refreshment samples, then $A(\kappa) = \prod_{i=n+1}^{m} f(y_i \mid \kappa)$. If we have an estimate $\widehat{\kappa}$ coming from a survey, then $A(\kappa) = f(\widehat{\kappa} \mid \kappa)$ is the density function from the approximate distribution of $\widehat{\kappa}$, such as the normal distribution in the case of Horvitz–Thompson estimators with large samples (e.g., Särndal et al., 1992, Chapter 2).

5 Illustrative example

5.1 Description of data and models

Each year in the U.S.A., the Behavioral Risk Factor Surveillance System collects data on risk factors associated with a variety of diseases. The data come from a random sample of adults contacted through a telephone survey (Centers for Disease Control and Prevention, 2010). We use the 2010 data to estimate the prevalence of diabetes among demographic strata formed by combinations of age, race, and sex. We focus on the U.S. Virgin Islands, as this territory has the highest nonresponse rates for 2010 in the variables that we study. Our study variables include diabetes, defined as $X \in \{\text{NO}, \text{YES}\}$, with 0.17% of nonresponse; age, defined as $Y_1 \in \{20\text{–}34, 35\text{–}49, 50\text{–}64, 65+\}$, with 2.47% of nonresponse; race, defined as $Y_2 \in \{\text{BLACK}, \text{WHITE}, \text{OTHER}\}$, with 5.89% of nonresponse; and sex, defined as $Y_3 \in \{\text{MALE}, \text{FEMALE}\}$, being fully observed. This ordering comes from the sequence in which the variables are recorded in the survey. The joint distribution of age, race and sex in the U.S. Virgin Islands is available from the 2010 decennial census.

We model $(X \mid Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) \sim \text{Bernoulli}\{\theta(y_1, y_2, y_3)\}$, with $\theta(y_1, y_2, y_3)$ representing the prevalence of diabetes in the stratum (y_1, y_2, y_3), and we fix the marginal

distribution of (Y_1, Y_2, Y_3) at its census value. We take a Bayesian approach to estimation and place flat priors on each $\theta(y_1, y_2, y_3)$. The missingness mechanism for X conditionally on (X, Y, M) is assumed to be ignorable, as explained in Section 4.2. We explore the estimation of the different per-stratum prevalences $\theta(y_1, y_2, y_3)$ under six different submodels of a full sequentially additive nonignorable missingness mechanism for $(M \mid X, Y)$, summarized in Table 1.

The first model that we consider is the full logit sequentially additive nonignorable missingness mechanism, which for the age nonresponse uses $\text{logit} f(M_1 = 1 \mid x, y) = \alpha_1(x, y_2, y_3) + \beta_1(y_1, y_2, y_3)$, with the values of α_1 for each (x, y_2, y_3) and β_1 for each (y_1, y_2, y_3) being the parameters of the model. This is equivalent to specifying the model in terms of a linear predictor using indicator variables for each (x, y_2, y_3) and for each (y_1, y_2, y_3). With the constraint $\beta_1(Y_1 = 20\text{–}34, y_2, y_3) = 0$, we obtain

$$\alpha_1(x, y_2, y_3) = \text{logit } f(M_1 = 1 \mid x, Y_1 = 20\text{–}34, y_2, y_3),$$

and for $y_1 \neq 20\text{–}34$,

$$\beta_1(y_1, y_2, y_3) = \log \frac{f(M_1 = 1 \mid x, Y_1 = y_1, y_2, y_3)/f(M_1 = 0 \mid x, Y_1 = y_1, y_2, y_3)}{f(M_1 = 1 \mid x, Y_1 = 20\text{–}34, y_2, y_3)/f(M_1 = 0 \mid x, Y_1 = 20\text{–}34, y_2, y_3)}.$$

This log-odds ratio explicitly captures the dependence of the nonresponse for the age variable on the categories of age. This dependence is constant across the diabetes status, but it can vary with race and sex. For the race question, the full logit sequentially additive nonignorable missingness mechanism given $M_1 = m_1$ uses

$$\text{logit } f(M_2 = 1 \mid x, y, m_1) = \text{logit } f(M_2 = 1 \mid x, y_1^*, y_{\geq 2}) = \alpha_2(x, y_1^*, y_3) + \beta_2(y_2, y_3). \quad (9)$$

With the constraint $\beta_2(\text{BLACK}, y_3) = 0$, we have

$$\alpha_2(x, y_1^*, y_3) = \text{logit } f(M_2 = 1 \mid x, y_1^*, Y_2 = \text{BLACK}, y_3),$$

and for $y_2 \neq \text{BLACK}$,

$$\beta_2(y_2, y_3) = \log \frac{f(M_2 = 1 \mid x, y_1^*, Y_2 = y_2, y_3)/f(M_2 = 0 \mid x, y_1^*, Y_2 = y_2, y_3)}{f(M_2 = 1 \mid x, y_1^*, Y_2 = \text{BLACK}, y_3)/f(M_2 = 0 \mid x, y_1^*, Y_2 = \text{BLACK}, y_3)}.$$

This log-odds ratio measures the association between the race variable and its nonresponse, which might be different across sex. We use independent normals with mean zero and standard deviation 1.5 as priors for each $\alpha_1(x, y_2, y_3)$ and each $\alpha_2(x, y_1^*, y_3)$, which lead to only slightly informative priors on the probability scale, and independent normals with mean zero and standard deviation 3 for each $\beta_1(y_1, y_2, y_3)$ and each $\beta_2(y_2, y_3)$, which are relatively spread out on the logit scale.

We can give plausible interpretations to the components of this comprehensive model. In particular, we now describe a scenario where the propensity to respond to the race question Y_2 depends on the materialized variable for age Y_1^*, as assumed in (9). Holding (X, Y_3) constant, consider two groups of people, those who do and do not respond to the age question. These groups could have different propensities to respond to the race question. For example, the first group could mostly include people who are willing to provide information to government agencies and hence are likely to respond to the race question, while the second group could mostly include people who believe that neither their age nor race—potentially sensitive variables—are the government's business and hence are unlikely to respond to the race question. For this second group, the individuals may decide whether or not to respond to the race question independent of their actual age; for example, everyone in this group may be sufficiently distrustful or disinterested in the survey so as not to respond to the sensitive questions. On the other hand, in the first group with generally response-compliant participants, it may be that younger people are less likely to report their race values than older people. For example, the younger people may feel that the categories of race listed on the survey do not describe their actual race, making them less likely to answer the question. Or, these younger participants may be more likely than older participants to believe that race, but not age, is a private matter, and hence less likely to respond than older people. As discussed in Section 3.3, the comprehensive version of the sequentially additive nonignorable missingness mechanism includes various ignorable and nonignorable models as special cases. Thus, using the comprehensive version is appropriate even if a plausible missingness mechanism is actually represented by a submodel.

In Table 1 we summarize five submodels of the comprehensive version presented above. The first submodel only has main effects, but still allows the nonresponse to directly depend on each item. The second and third submodels are invariant to the ordering of the variables, with the third one representing a mechanism where the nonresponse directly depends only on the variable that we attempt to measure. In the fourth submodel the nonresponse for each item does not directly depend on the item itself, but it is still nonignorable since $f(M_1 = 1 \mid x, y)$ depends on the unobserved y_2 values. The fifth submodel corresponds to an ignorable mechanism. The terms in each of these submodels have analogous interpretations as those in the full model, and therefore we similarly impose normal priors with mean zero and standard deviation 1.5 on each α-term, and normals with mean zero and standard deviation 3 for the non-zero β-terms.

We obtained approximate posterior distributions for the parameters of these six models using a standard Gibbs sampler, with a data augmentation scheme for the missing data (Tanner and Wong, 1987), and the strategy of Polson et al. (2013) to expand the parts of the likelihood functions coming from the logistic regressions in terms of Pólya–Gamma latent variables.

5.2 Results

In Fig. 1 we present the posterior distributions of $\theta(y_1, y_2, \text{FEMALE})$, that is, the diabetes prevalence among females of age y_1 and race y_2, under the six different missingness mechanisms of Table 1. The posteriors under models 0–3 are very similar, all of which encode a direct dependence of the nonresponse on the corresponding study variables. On the other hand, models 4 and 5, which exclude a direct dependence of the missingness mechanism on the study variables, also lead to nearly the same posterior distributions. This indicates that the differences obtained between assuming ignorability and sequential additive nonignorability are mainly due to the direct dependence of the nonresponse on the study variables. Namely, for these data the most relevant feature of the sequentially additive nonignorable mechamism is represented by submodel 3. Figures 2 and 3 make this

Table 1: Subclasses of sequential additive nonignorability explored in Section 5.

	$\text{logit} f(M_1 = 1 \mid x, y)$	$\text{logit} f(M_2 = 1 \mid x, y_1^*, y_{\geq 2})$
0. Full	$\alpha_1(x, y_2, y_3) + \beta_1(y_1, y_2, y_3)$	$\alpha_2(x, y_1^*, y_3) + \beta_2(y_2, y_3)$
Restrictions	$\beta_1(Y_1 = 20\text{–}34, y_2, y_3) = 0$	$\beta_2(Y_2 = \text{BLACK}, y_3) = 0$
1. Main effects	$\alpha_1(x) + \sum_{j=1}^{3} \beta_{1j}(y_j)$	$\alpha_2(x) + \beta_{21}(y_1^*) + \sum_{j=2}^{3} \beta_{2j}(y_j)$
Restrictions	$\beta_{11}(20\text{–}34) = \beta_{12}(\text{BLACK}) = \beta_{13}(\text{MALE}) = 0$	$\beta_{21}(*) = \beta_{22}(\text{BLACK}) = \beta_{23}(\text{MALE}) = 0$
2. Order-invariant	$\alpha_1(x, y_3) + \beta_1(y_1, y_3)$	$\alpha_2(x, y_3) + \beta_2(y_2, y_3)$
Restrictions	$\beta_1(Y_1 = 20\text{–}34, y_3) = 0$	$\beta_2(Y_2 = \text{BLACK}, y_3) = 0$
3. Only-directly dependent	$\alpha_1 + \beta_1(y_1)$	$\alpha_2 + \beta_2(y_2)$
Restrictions	$\beta_1(Y_1 = 20\text{–}34) = 0$	$\beta_2(Y_2 = \text{BLACK}) = 0$
4. Not-directly dependent	$\alpha_1(x, y_2, y_3)$	$\alpha_2(x, y_1^*, y_3)$
5. Ignorable	$\alpha_1(x)$	$\alpha_2(x, y_1^*)$

evident, as we explain below. The posterior distributions of $\theta(y_1, y_2, \text{MALE})$ were not very sensitive to the missingness mechanism so we omit them.

In Fig. 2 we present the posterior distributions of the log-odds ratios $\beta_1(65+, y_2, y_3)$ in the full model for race y_2 and sex y_3. The analogous posteriors for the age categories 35–49 and 50–64 are very similar, so we omit them. Most of the mass of each of these posteriors is below zero, indicating that in this illustration the odds of nonresponse in the variable age when the respondent is 65+ are likely to be much smaller than when he or she is 20–34, which is the baseline. This indicates a strong negative association between the variable age and its nonresponse.

In Fig. 3 we present the posterior distributions of the log-odds ratios $\beta_2(y_2, y_3)$ in the full model for race y_2 and sex y_3. The posteriors for both $\beta_2(\text{WHITE}, \text{MALE})$ and $\beta_2(\text{WHITE}, \text{FEMALE})$ have masses mostly below zero, indicating that in this illustration the odds of nonresponse in the variable race for white respondents are likely to be much smaller

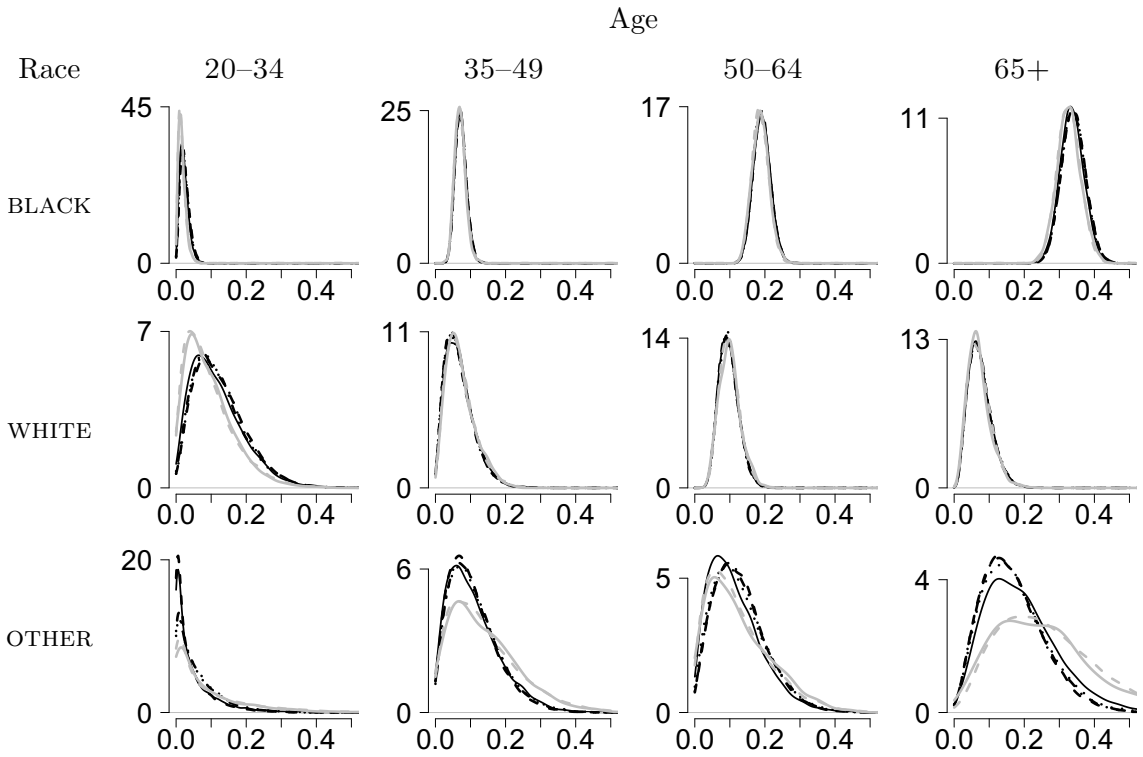

Figure 1: Posterior distributions of proportion of females with diabetes among combinations of age and race under different missingness mechanisms. Model 0: full sequential additive nonignorability, black solid line; model 1: main effects, black dashed line; model 2: order-invariant, black dot-dashed line; model 3: only-directly dependent, black dotted line; model 4: not-directly dependent, gray dashed line; model 5: ignorable model, gray solid line.

than for black respondents. The posterior of $\beta_2(\text{OTHER}, \text{MALE})$ is centered around zero, indicating that the odds of nonresponse in the race variable among males is similar for blacks and for people of other race. On the other hand, the posterior of $\beta_2(\text{OTHER}, \text{FEMALE})$ is fully concentrated above zero, indicating that the odds of nonresponse in the race variable are higher for females of other race compared with black females.

As mentioned throughout the article, the assumptions encoded by the sequentially additive nonignorable missingness mechanism rely on an ordering of the variables, and changing this order leads to different missingness mechanisms, with their full versions enjoying

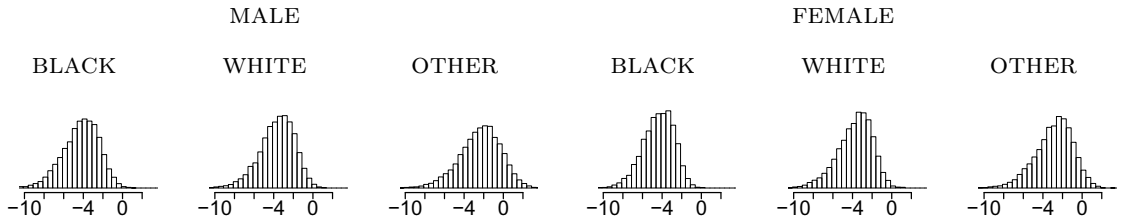

Figure 2: Posterior distributions of log-odds ratios of nonresponse in age question for age 65+ versus baseline 20–34, for combinations of race and sex.

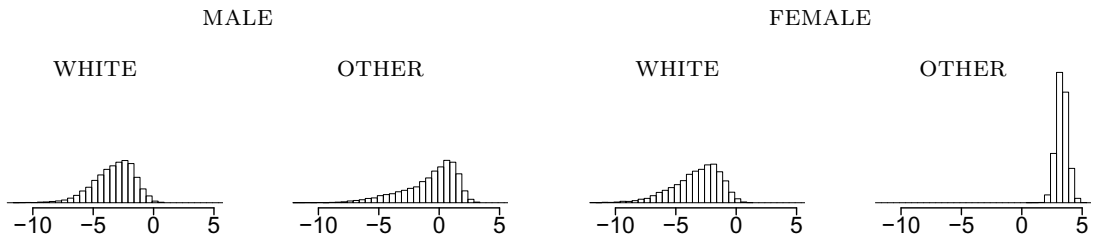

Figure 3: Posterior distributions of log-odds ratios of nonresponse in race question for races WHITE and OTHER versus baseline BLACK, among males and females.

nonparametric identification. This property leads to a natural way of performing global or local sensitivity analyses (e.g., Scharfstein et al., 2018) by obtaining inferences under completely or partially different orderings of the variables, respectively.

The main analysis that we presented relied on the order in which the variables were collected, but we also performed our analyses changing the order of the age and race variables. The order of the sex variable is not relevant as it is fully observed, nor is the order of the diabetes variable since its missingness is assumed to be ignorable. By construction, models 2, 3 and 5 do not depend on the order of age and race. Changing the order of age and race in models 0, 1 and 4 could potentially lead to very different results, but with these data we found that the posterior distributions of the proportions of people with diabetes were virtually the same as in Fig. 1. This finding is intuitive given that the results in Fig. 1 are essentially the same for models 0–3, and models 2 and 3 do not depend on the ordering of age and race. We also replicated all of our results using the probit link, using the data

augmentation scheme of Albert and Chib (1993), and found that the results were virtually the same as using the logit link.

The results of this illustration indicate the existence of direct dependence of the nonresponse on the values of the items or variables being measured. This direct dependence can be quantified thanks to the availability of auxiliary marginal information, in this case coming from a census, and thanks to the theoretical results presented in this article, which permits us to identify missingness mechanisms where this direct dependence occurs.

6 Final remarks and future work

The implementation that we proposed in Section 4.3 relied on a likelihood function where the density of the distribution of Y, $f(y \mid \kappa)$, is parameterized explicitly in terms of features for which we have auxiliary marginal information. This is not restrictive in the case of categorical variables, as in our application, but it might become so for continuous variables. For example, if one is willing to assume that Y is multivariate Gaussian, κ would represent means, variances and covariances. We could then use auxiliary marginal information on some of those parameters, but information beyond the first two moments could not be used in this case. Our identification results, however, are general enough to allow an arbitrary modeling of the distribution of Y. Thus, a natural future avenue of research is to combine nonparametric or highly flexible models of the distribution of Y with sequentially additive nonignorable missingness mechanisms. For example, we could incorporate information on summaries of distributions into nonparametric Bayesian models using the approach of Kessler et al. (2015).

Acknowledgement

The authors were supported by two grants from the U.S.A. National Science Foundation.

Appendix 1

$\mathbb{f}$-projections

The monograph of Liese and Vajda (1987) contains a compendium of results on $\mathbb{f}$-projections, some of which we present here for completeness. We let P and Q be two probability distributions, $\mathcal{P}$ be a set of probability distributions, and $I_{\mathbb{f}}(P,Q) = \int \mathbb{f}(dP/dQ)dQ$ be the $\mathbb{f}$-divergence of P and Q. We start with Propositions 8.2 and 8.5 of Liese and Vajda (1987) which address the uniqueness and existence of $\mathbb{f}$-projections.

Theorem 5 (Uniqueness of $\mathbb{f}$-projections). *Let $\mathcal{P}$ be convex, where each $P \in \mathcal{P}$ is dominated by Q. Let $\mathbb{f}$ be strictly convex at every point of $(0,\infty)$ and let $\inf_{P\in\mathcal{P}} I_{\mathbb{f}}(P,Q) < \infty$. Then there exists at most one $\mathbb{f}$-projection of Q onto $\mathcal{P}$.*

Theorem 6 (Existence of $\mathbb{f}$-projections). *Let $\mathcal{P}$ be convex and closed in variational distance. Let $\lim_{z\to\infty} \mathbb{f}(z)/z = \infty$. Then there exists an $\mathbb{f}$-projection of Q onto $\mathcal{P}$.*

The following result is simply a version of Theorem 8.20 in Liese and Vajda (1987) after combining it with their Lemma 8.19.

Theorem 7 (Characterization of $\mathbb{f}$-projections). *Let $\mathbb{f}$ be a strictly convex differentiable function that satisfies Assumption 2. Let $\mathbb{f}^*(z) = z\mathbb{f}(1/z)$ for $z \in (0,\infty)$ also satisfy Assumption 2. Let $\mathcal{P}(\mathcal{U})$ be the set of probability distributions where each $P \in \mathcal{P}(\mathcal{U})$ has the same known finite value of $\int u dP$ for each function u in a set $\mathcal{U}$. Then $\langle\mathcal{U}\rangle \subset L_1(P)$ for each $P \in \mathcal{P}(\mathcal{U})$ and*

1. *$P^* \in \mathcal{P}(\mathcal{U})$ is the $\mathbb{f}$-projection of Q onto $\mathcal{P}(\mathcal{U})$ if $\mathbb{f}'(dP^*/dQ) \in \langle\mathcal{U}\rangle$.*
2. *If P^* is the $\mathbb{f}$-projection of Q onto $\mathcal{P}(\mathcal{U})$ then $\mathbb{f}'(dP^*/dQ) \in \overline{\langle\mathcal{U}\rangle}$, where the closure is in $L_1(P^*)$.*

Similar results to this characterization can be found in Csiszár (1975), Rüschendorf (1984), and Broniatowski and Keziou (2006).

Appendix 2

Proofs

Proof [of Theorem 1] The $\mathfrak{f}_\lambda$ function is differentiable by construction, and it is strictly convex given that its derivative $\mathfrak{f}'_\lambda(z) = \lambda[z/(c+z)]$ is monotonically increasing in $(0, \infty)$. We can see that $\lim_{z\to\infty} \mathfrak{f}^*_\lambda(z)/z = 0$, which is a sufficient condition for $\mathfrak{f}^*_\lambda$ to satisfy Assumption 2 (Liese and Vajda, 1987, p. 171). Now, we can rewrite $f(x, y \mid M = 1) = f(x, y \mid M = 0)\varphi\{f(M = 1 \mid x, y)\}$, where $\varphi(z) = cz/(1 - z)$, with $c = (1 - \pi)/\pi$. We also assumed $f(M = 1 \mid x, y) = \lambda^{-1}\{\alpha(x)+\beta(y)\}$. This means that $dF_1/dF_0(x, y) = \varphi[\lambda^{-1}\{\alpha(x)+\beta(y)\}]$, where F_m denotes the distribution of $X, Y \mid M = m$, for $m = 0, 1$. We then obtain $\mathfrak{f}'_\lambda\{dF_1/dF_0(x, y)\} = \alpha(x) + \beta(y) \in \langle L_1\{f(x \mid M = 1)\} \cup \mathcal{U}\rangle$. Theorem 7 then implies that F_1 is the $\mathfrak{f}_\lambda$-projection of F_0 onto the set of distributions with X-marginal given by $f(x \mid M = 1)$ and with expected values of each $u \in \mathcal{U}$ matching those determined by the auxiliary marginal information, $\{\int u(y)f(y \mid M = 1)\mu(dy)\}_{u\in\mathcal{U}}$. Theorem 1 simply states this result in terms of densities of F_1 and F_0. □

Proof [of Theorem 2] Let us denote by G_m and H_m the distributions with densities $g(x, y \mid M = m)$ and $h(x, y \mid M = m)$, respectively, $m = 0, 1$. Let us denote by P^X and P^Y the X- and Y-marginals of a distribution P. By assumption, G_1 is the $\mathfrak{f}_\lambda$-projection of H_0 onto the set of distributions $\mathcal{P}$ where each element P is such that $\int v dP^X = \int v dH_1^X$ and $\int u dP^Y = \int u dH_1^Y$, for all $v \in L_1(H_1^X)$ and all $u \in \mathcal{U}$. To guarantee the existence of G_1, we note that $\mathcal{P}$ is convex as convex combinations of distributions that satisfy the constraints also satisfy the constraints, and is closed in variational distance since it is the solution set of a number of equations given by the constraints. Furthermore, by L'Hôpital's rule we find that $\lim_{z\to\infty} \mathfrak{f}_\lambda(z)/z = \lim_{z\to\infty} \lambda\{z/(c + z)\} = \infty$. Therefore Theorem 6 guarantees the existence of G_1, and its uniqueness is obtained from Theorem 5 since $\mathfrak{f}_\lambda$ is strictly convex given that its derivative $\mathfrak{f}'_\lambda(z) = \lambda[z/(c + z)]$ is monotonically increasing in $(0, \infty)$.

Similarly as in the proof of Theorem 1, we find that $\mathfrak{f}_\lambda$ and $\mathfrak{f}^*_\lambda$ satisfy the conditions of Theorem 7. Therefore Theorem 7 implies that $\mathfrak{f}'_\lambda(dG_1/dH_0) \in \overline{\langle L_1(H_1^X) \cup \mathcal{U}\rangle}$, where the

functions in $L_1(H_1^X)$ are constant in y and the functions in $\mathcal{U}$ are constant in x. Note that if $\mathfrak{f}'_\lambda(dG_1/dH_0) \in \langle L_1(H_1^X) \cup \mathcal{U}\rangle$, then it can be written as $\mathfrak{f}'_\lambda(dG_1/dH_0) = \alpha(x) + \beta(y)$ where $\alpha \in L_1\{H_1^X\}$ and $\beta \in \langle \mathcal{U}\rangle$, but if it is a limit point outside of that set then it can be arbitrarily approximated by functions of such form. Now, from the definition of $g(x, y, m)$, we find that $g(M = 1 \mid x, y) = dG_1/dH_0(x, y)/\{c + dG_1/dH_0(x, y)\}$, since $g(x, y \mid M = 1)/h(x, y \mid M = 0) = dG_1/dH_0(x, y)$. Therefore, we conclude $\lambda[g(M = 1 \mid x, y)] = \mathfrak{f}'_\lambda[dG_1/dH_0(x, y)] \in \overline{\langle L_1(H_1^X) \cup \mathcal{U}\rangle}$. Now, $g(x, y^*) = h(x, y^*)$ because by construction $g(x, y, M = 0) = h(x, y, M = 0)$, and $g(x, M = 1) = \pi g(x \mid M = 1) = \pi h(x \mid M = 1)$ from how we construct $g(x, y \mid M = 1)$ as an $\mathfrak{f}_\lambda$-projection. Finally, we also have that $\int u(y)g(y)\mu(dy) = \int u(y)h(y)\mu(dy)$ because $g(M = 1) = h(M = 1)$, and $\int u(y)g(y \mid M = m)\mu(dy) = \int u(y)h(y \mid M = m)\mu(dy)$ for all $u \in \mathcal{U}$, when $m = 1$ based on how we construct $g(x, y \mid M = 1)$ as an $\mathfrak{f}_\lambda$-projection, and when $m = 0$ given that $g(y \mid M = 0) = h(y \mid M = 0)$ by construction of $g(x, y, m)$. □

Proof [of Theorem 3] 1. Analogously to the proof of Theorem 1, we first find that all $\mathfrak{f}_{\lambda,j}$ and $\mathfrak{f}^*_{\lambda,j}$ satisfy the conditions required by Theorem 7, and we obtain $\mathfrak{f}'_{\lambda,j}\{dF_{1,j}/dF_{0,j}(x, y)\} = \alpha_j(x, y^*_{<j}, y_{>j}) + \beta_j(y_{\geq j}) \in \langle L_1\{f(x, y^*_{<j}, y_{>j} \mid M = 1)\} \cup \mathcal{U}_{\geq j}\rangle$, where $F_{m,j}$ is the distribution with density $f(x, y^*_{<j}, y_{\geq j} \mid M_j = m)$, $m = 0, 1$. Theorem 7 then implies that $F_{1,j}$ is the $\mathfrak{f}_{\lambda,j}$-projection of $F_{0,j}$ onto the set of distributions with marginal determined by $f(x, y^*_{<j}, y_{>j} \mid M = 1)$ and with expected values of each $u \in \mathcal{U}_{\geq j}$ matching those determined by the auxiliary marginal information, $\{E[u(Y_{\geq j}) \mid M_j = 1]; u \in \mathcal{U}_{\geq j}\}$.

2. Part 1 of this theorem guarantees that the true density $f(x, y^*_{<j}, y_{\geq j} \mid M_j = 1)$ can be recovered from $f(x, y^*_{<j}, y_{\geq j} \mid M_j = 0)$, $f(x, y^*_{<j}, y_{>j} \mid M_j = 1)$ and $\{E[u(Y_{\geq j}) \mid M_j = 1]; u \in \mathcal{U}_{\geq j}\}$, for each $j = 1, \ldots, p$. Algorithm 1 implements the sequence of projections justified by Part 1. For $j = 1$ we obtain $f(x, y)$ from the algorithm's substep d, and the missingness mechanism is obtained as $f(m \mid x, y) = \prod_{j=1}^p f(m_j \mid x, y^*_{<j}, y_{\geq j})$, where $f(m_j \mid x, y^*_{<j}, y_{\geq j})$ is obtained in step j of the algorithm. □

Proof [of Theorem 4] 1. In the construction of $g(x, y, m)$, we find $g(x, y^*_{<j}, y_{\geq j} \mid M_j = 1)$ as the $\mathfrak{f}_{\lambda,j}$-projection of $g(x, y^*_{<j}, y_{\geq j} \mid M_j = 0)$ onto the set of distributions that match

the marginal $g(x, y^*_{<j}, y_{>j} \mid M_j = 1)$ and the expectations $\{E_g[u(Y_{\geq j}) \mid M_j = 1]; u \in \mathcal{U}_{\geq j}\}$, with $\mathfrak{f}_{\lambda,j}(z) = \int_0^z \lambda[v/(c_j + v)]dv$, $c_j = (1 - \pi_j)/\pi_j$. Confirming the conditions required by Theorems 6 and 5 to guarantee the existence and uniqueness of each $g(x, y^*_{<j}, y_{\geq j} \mid M_j = 1)$ is analogous as in the proof of Theorem 2, so we omit it. Finding that $\mathfrak{f}_{\lambda,j}$ and $\mathfrak{f}^*_{\lambda,j}$ satisfy the conditions of Theorem 7 is also analogous as in the proof of Theorem 1, so we also omit it. Denoting by $G_{m,j}$ the distribution with density $g(x, y^*_{<j}, y_{\geq j} \mid M_j = m)$, $m = 0, 1$, Theorem 7 implies that $\mathfrak{f}'_{\lambda,j}(dG_{1,j}/dG_{0,j}) \in \overline{\langle L_1\{g(x, y^*_{<j}, y_{>j} \mid M_j = 1)\} \cup \mathcal{U}_{\geq j}\rangle}$, where the functions in $L_1\{g(x, y^*_{<j}, y_{>j} \mid M_j = 1)\}$ are constant in y_j and the functions in $\mathcal{U}_{\geq j}$ are constant in $(x, y^*_{<j})$. Now, by construction we find that $g(M_j = 1 \mid x, y^*_{<j}, y_{\geq j}) = dG_{1,j}/dG_{0,j}(x, y^*_{<j}, y_{\geq j})/\{c_j + dG_{1,j}/dG_{0,j}(x, y^*_{<j}, y_{\geq j})\}$, since $g(x, y^*_{<j}, y_{\geq j} \mid M_j = 1)/g(x, y^*_{<j}, y_{\geq j} \mid M_j = 0) = dG_{1,j}/dG_{0,j}(x, y)$. Therefore, we conclude $\lambda[g(M_j = 1 \mid x, y^*_{<j}, y_{\geq j})] = \mathfrak{f}'_{\lambda,j}[dG_{1,j}/dG_{0,j}(x, y^*_{<j}, y_{\geq j})] \in \overline{\langle L_1\{g(x, y^*_{<j}, y_{>j} \mid M_j = 1)\} \cup \mathcal{U}_{\geq j}\rangle}$.

2. We first show that the observed-data distribution implied by $g(x, y, m)$ is $h(x, y^*)$. The result of Algorithm 1 is $g(x, y, m) = g(x, y) \prod_{j=1}^p g(m_j \mid x, y^*_{<j}, y_{\geq j})$, which we need to integrate over the missing values y_m according to a generic missingness pattern $m = (m_1, \ldots, m_p)$. To do this we sequentially integrate $g(x, y, m)$ over y_j if $m_j = 1$, $j = 1, \ldots, p$. At each step we obtain $g(x, y^*_{\leq j}, y_{>j}, m_{>j})$ from $g(x, y^*_{<j}, y_{\geq j}, m_{\geq j})$. When $j = 1$ this corresponds to obtaining $g(x, y^*_1, y_{>1}, m_{>1})$ from $g(x, y, m)$, and when $j = p$ this corresponds to obtaining $g(x, y^*)$ from $g(x, y^*_{<p}, y_p, m_p)$.

Let us say that after having integrated over $(y_l : m_l = 1, l < j)$ we obtained $g(x, y^*_{<j}, y_{\geq j}, m_{\geq j})$. If $m_j = 0$ then we do not have to integrate over y_j, and $g(x, y^*_{<j}, y_{\geq j}, m_{\geq j}) = g(x, y^*_{\leq j}, y_{>j}, m_{>j})$ from the definition of Y^*_j. If $m_j = 1$ then we need to integrate over y_j. From the construction in Algorithm 1, $g(x, y^*_{<j}, y_{\geq j}, m_{\geq j}) = g(x, y^*_{<j}, y_{\geq j}) \prod_{k \geq j} g(m_k \mid x, y^*_{<k}, y_{\geq k})$. By definition, all $g(m_k \mid x, y^*_{<k}, y_{\geq k})$ for $k > j$ do not depend on y_j when $m_j = 1$, and so we just need to integrate $g(x, y^*_{<j}, y_{\geq j})g(M_j = 1 \mid x, y^*_{<j}, y_{\geq j}) = g(x, y^*_{<j}, y_{\geq j} \mid M_j = 1)\pi_j$ over y_j. Now, $g(x, y^*_{<j}, y_{\geq j} \mid M_j = 1)$ is obtained as the $\mathfrak{f}_{\lambda,j}$-projection of $g(x, y^*_{<j}, y_{\geq j} \mid M_j = 0)$ onto the set constrained by the marginal $g(x, y^*_{<j}, y_{>j} \mid M_j = 1)$ and the expectations $\{E_g[u(Y_{\geq j}) \mid M_j = 1]; u \in \mathcal{U}_{\geq j}\}$. By construction then $\int g(x, y^*_{<j}, y_{\geq j} \mid M_j =$

$1)\mu_j(dy_j)\pi_j = g(x, y^*_{<j}, y_{>j} \mid M_j = 1)\pi_j$, which can be written as $g(x, y^*_{\leq j}, y_{>j})$ from the definition of Y^*_j when $M_j = 1$. We then obtain for $m_j = 1$, $g(x, y^*_{\leq j}, y_{>j}, m_{>j}) = \int g(x, y^*_{<j}, y_{\geq j}, m_{\geq j})\mu_j(dy_j) = g(x, y^*_{\leq j}, y_{>j}) \prod_{k>j} g(m_k \mid x, y^*_{<k}, y_{\geq k})$. When $j = p$ we then obtain $g(x, y^*_{\leq p}, y_{>p}, m_{>p}) = g(x, y^*) = h(x, y^*)$.

Finally, to show that $\int u(y)g(y)\mu(dy) = \int u(y)h(y)\mu(dy)$ for all $u \in \mathcal{U}$, note that from the output of Algorithm 1 $g(x, y) = g(x, y \mid M_1 = 1)\pi_1 + g(x, y \mid M_1 = 0)(1 - \pi_1)$ from substep d. of the algorithm's last step. Here $g(x, y \mid M_1 = 1)$ is the $\mathbb{f}_{\lambda,1}$-projection of $g(x, y \mid M_1 = 0)$ onto the set with marginal given by $g(x, y_{>1} \mid M_1 = 1)$ and expectations given by $\{E_g[u(Y) \mid M_1 = 1]; u \in \mathcal{U}\}$, as derived from steps a. and b. of the algorithm's last step. We therefore obtain that $\int u(y)g(y)\mu(dy) = \int u(y)g(y \mid M_1 = 1)\mu(dy)\pi_1 + \int u(y)g(y \mid M_1 = 0)\mu(dy)(1-\pi_1) = \int u(y)h(y)\mu(dy)$, given that $\int u(y)g(y \mid M_1 = 1)\mu(dy) = E_g[u(Y) \mid M_1 = 1] = \{E[u(Y)] - \int u(y)g(y, M_1 = 0)\mu(dy)\}/\pi_1$. □

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Assoc.*, 88(422):669–679.

Aranda-Ordaz, F. J. (1981). On Two Families of Transformations to Additivity for Binary Response Data. *Biometrika*, 68(2):357–363.

Berrocal, V. J., Miranda, M. L., Gelfand, A. E., and Bhattacharya, S. (2013). Synthesizing categorical datasets to enhance inference. *Statist. Method.*, 15:25–45.

Bhattacharya, B. (2006). An iterative procedure for general probability measures to obtain I-projections onto intersections of convex sets. *Ann. Statist.*, 34(2):878–902.

Bhattacharya, D. (2008). Inference in panel data models under attrition caused by unobservables. *J. Econometrics*, 144(2):430 – 446.

Broniatowski, M. and Keziou, A. (2006). Minimization of divergences on sets of signed measures. *Studia Scientiarum Mathematicarum Hungarica*, 43(4):403–442.

Centers for Disease Control and Prevention (2010). Behavioral Risk Factor Surveillance System Survey Data. Atlanta, Georgia: U.S. Department of Health and Human Services, available at https://www.cdc.gov/brfss/.

Chatterjee, N., Chen, Y.-H., Maas, P., and Carroll, R. J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *J. Am. Statist. Assoc.*, 111(513):107–117.

Choudhuri, N., Ghosal, S., and Roy, A. (2007). Nonparametric binary regression using a gaussian process prior. *Statist. Method.*, 4(2):227–243.

Csiszár, I. (1963). Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *Magyar. Tud. Akad. Mat. Kutató Int. Közl*, 8:85–108.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, 3(1):146–158.

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.

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(2):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*, pages 255–294.

Greenlees, J. S., Reece, W. S., and Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *J. Am. Statist. Assoc.*, 77(378):251–261.

Guo, Y., Little, R., and McConnell, D. S. (2012). On using summary statistics from an external calibration sample to correct for measurement error. *Epidemiology*, 23:165–174.

Harel, O. and Schafer, J. L. (2009). Partial and latent ignorability in missing-data problems. *Biometrika*, 96(1):37–50.

Hausman, J. and Wise, D. (1979). Attrition bias in experimental and panel data: the Gary income maintenance experiment. *Econometrica*, 47:455–473.

Hirano, K., Imbens, G. W., Ridder, G., and Rubin, D. B. (1998). Combining Panel Data Sets with Attrition and Refreshment Samples. Technical Report 230, National Bureau of Economic Research.

Hirano, K., Imbens, G. W., Ridder, G., and Rubin, D. B. (2001). Combining panel data sets with attrition and refreshment samples. *Econometrica*, 69(6):1645–1659.

Hoonhout, P. and Ridder, G. (2018). Nonignorable Attrition in Multi-Period Panels With Refreshment Samples. *J. Bus. Econ. Statist.*, Forthcoming.

Ibrahim, J. G., Lipsitz, S. R., and Chen, M. H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *J. R. Statist. Soc.* B, 61:173–190.

Kessler, D. C., Hoff, P. D., and Dunson, D. B. (2015). Marginally specified priors for non-parametric Bayesian estimation. *J. R. Statist. Soc.* B, 77(1):35–58.

Koopmans, T. C. (1949). Identification problems in economic model construction. *Econometrica*, 17(2):125–144.

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*. Springer, New York, 2nd edition.

Liese, F. and Vajda, I. (1987). *Convex Statistical Distances*. Teubner-Texte zur Mathematik, Leipzig.

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

Lohr, S. L. (2010). *Sampling: Design and Analysis, Second Edition*. New York: Cengage Learning.

Mealli, F. and Rubin, D. B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika*, 102(4):995–1000.

National Academies of Sciences, Engineering, and Medicine (2017). *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington D.C.: National Academies Press.

Nevo, A. (2003). Using weights to adjust for sample selection when auxiliary information is available. *J. Bus. Econ. Statist.*, 21(1):43–52.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *J. Am. Statist. Assoc.*, 108(504):1339–1349.

Robins, J. M. (1997). Non-response models for the analysis of non-monotone non-ignorable missing data. *Statist. Med.*, 16(1):21–37.

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

Rüschendorf, L. (1984). On the minimum discrimination information theorem. *Statistics & Decisions*, Supplement Issue No. 1:263–283.

Rüschendorf, L. (1995). Convergence of the iterative proportional fitting procedure. *Ann. Statist.*, 23(4):1160–1174.

Sadinle, M. and Reiter, J. P. (2017). Itemwise conditionally independent nonresponse modelling for incomplete multivariate data. *Biometrika*, 104(1):207–220.

Sadinle, M. and Reiter, J. P. (2018). Sequential identification of nonignorable missing data mechanisms. *Statist. Sinica*, 28(4):1741–1759.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Scharfstein, D., McDermott, A., Díaz, I., Carone, M., Lunardon, N., and Turkoz, I. (2018).

Global sensitivity analysis for repeated measures studies with informative drop-out: A semi-parametric approach. *Biometrics*, 74(1):207–219.

Seaman, S., Galati, J., Jackson, D., and Carlin, J. (2013). What is meant by "missing at random"? *Statist. Sci.*, 28(2):257–268.

Si, Y., Reiter, J. P., and Hillygus, D. S. (2015). Semi-parametric selection models for potentially non-ignorable attrition in panel studies with refreshment samples. *Political Analysis*, 23(1):92–112.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Am. Statist. Assoc.*, 82(398):528–540.

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(3):953–979.

Xu, D., Daniels, M. J., and Winterstein, A. G. (2016). Sequential BART for imputation of missing covariates. *Biostatistics*, 17:589–602.