Statistica Sinica **22** (2012), 149-172
doi:http://dx.doi.org/10.5705/ss.2010.069

DOUBLY ROBUST NONPARAMETRIC MULTIPLE IMPUTATION FOR IGNORABLE MISSING DATA

Qi Long, Chiu-Hsieh Hsu and Yisheng Li

Emory University, University of Arizona and
University of Texas MD Anderson Cancer Center

Abstract: Missing data are common in medical and social science studies and often pose a serious challenge in data analysis. Multiple imputation methods are popular and natural tools for handling missing data, replacing each missing value with a set of plausible values that represent the uncertainty about the underlying values. We consider a case of missing at random (MAR) and investigate the estimation of the marginal mean of an outcome variable in the presence of missing values when a set of fully observed covariates is available. We propose a new nonparametric multiple imputation (MI) approach that uses two working models to achieve dimension reduction and define the imputing sets for the missing observations. Compared with existing nonparametric imputation procedures, our approach can better handle covariates of high dimension, and is doubly robust in the sense that the resulting estimator remains consistent if either of the working models is correctly specified. Compared with existing doubly robust methods, our nonparametric MI approach is more robust to the misspecification of both working models; it also avoids the use of inverse-weighting and hence is less sensitive to missing probabilities that are close to 1. We propose a sensitivity analysis for evaluating the validity of the working models, allowing investigators to choose the optimal weights so that the resulting estimator relies either completely or more heavily on the working model that is likely to be correctly specified and achieves improved efficiency. We investigate the asymptotic properties of the proposed estimator, and perform simulation studies to show that the proposed method compares favorably with some existing methods in finite samples. The proposed method is further illustrated using data from a colorectal adenoma study.

Key words and phrases: Doubly robust, missing at random, multiple imputation, nearest neighbor, nonparametric imputation, sensitivity analysis.

1. Introduction

Missing data are common in medical and social science studies. When the missingness does not depend on the data values, missing or observed, the data are called missing completely at random (MCAR) and one can perform a so-called complete case analysis by ignoring the observations with missing values (Little and Rubin (2002)). In practice, the assumption of MCAR is rarely satisfied, and

a complete case analysis can lead to biased estimation. A more realistic missing mechanism is that the data are missing at random (MAR) (Rubin (1987); Little and Rubin (2002)), which means that the missing status is independent of missing values conditional on observed covariates.

1.1. The problem

Let Y denote the outcome variable of interest with missing values, δ denote the missingness indicator, $\delta = 0$ if Y is missing and $\delta = 1$ if Y is observed, and $\mathbf{X} = (X_1, \ldots, X_p)$ denote a set of fully observed covariates that are predictive of Y or δ. Suppose $(Y_i, \mathbf{X}_i, \delta_i)$ $(i = 1, \ldots, n)$ constitute an independent and identically distributed sample from n subjects. The observed data can be written as $(\delta_i Y_i, \mathbf{X}_i, \delta_i)$ $(i = 1, \ldots, n)$ where $\delta_i Y_i$ is missing when $\delta_i = 0$. We consider the estimation of $\mu = E(Y)$ when Y is independent of δ given $\mathbf{X}$. For this type of problems, one can use imputation methods (either single or multiple) or the inverse probability weighting methods that are doubly robust.

1.2. Imputation methods

An imputation method replaces each missing value with one "plausible" value (single imputation) or a set of "plausible" values (multiple imputation, MI), and subsequently standard analysis is performed using the imputed datasets. Adjustments are necessary for computing standard errors to account for the uncertainty of the imputed values (Rubin (1987); Little and Rubin (2002)). The imputation models can be parametric (Matloff (1981); Little and Rubin (2002)), semiparametric (Wang, Linton, and Härdle (2004)) or nonparametric (Titterington and Sedransk (1989); Cheng (1994); Aerts et al. (2002)). Despite being efficient when the parametric component is correctly specified, the parametric and semiparametric imputation methods are sensitive to model misspecifications. While a nonparametric imputation approach is more robust to model misspecification, a different challenge arises. Most existing nonparametric regression imputation methods focus on the case of a single fully observed covariate. For example, Cheng (1994) studied a single imputation approach that imputes missing values with kernel estimates of the conditional mean of an outcome variable given a continuous covariate; Aerts et al. (2002) studied an MI approach, which was based on the nonparametrically estimated distribution of the outcome variable conditional on the covariate using kernel methods, among others. The main difficulty with nonparametric imputation methods is that as the number of covariates increases, it becomes increasingly difficult to estimate either the conditional distribution or the conditional expectation of the outcome variable given the covariates, due to the curse of dimensionality. In practice, performance loss for the nonparametric

imputation methods can be substantial even when only a small number of covariates are used. It is important to have a nonparametric approach that alleviates the curse of dimensionality in the presence of multiple covariates.

1.3. Doubly robust methods

The earliest doubly robust method for missing data is the calibration estimator (CE), also known as the the generalized regression estimator (Cassel, Sarndal, and Wretman (1976)), which extended the inverse probability weighting method (Horvitz and Thompson (1952)). The CE uses two working models based on the observed covariates: one model predicts the missing values, and the other model predicts the missing probabilities. Specifically, the estimator is a result of expressing the mean of Y as a sum of prediction and inverse probability-weighted prediction errors,

$$\mu = E[E(Y|\mathbf{X})] + E[\delta \frac{(Y - E(Y|\mathbf{X}))}{\pi(\mathbf{X})}],$$

where $\pi(\mathbf{X}) = E(\delta|\mathbf{X})$. Plugging the estimate of each parameter into the expression leads to

$$\hat{\mu}_{CE} = n^{-1} \sum_{i=1}^{n} \hat{Y}_i + n^{-1} \sum_{i=1}^{n} \delta_i w_i (Y_i - \hat{Y}_i) = n^{-1} \sum_{i=1}^{n} \hat{Y}_i + n^{-1} \sum_{i=1}^{n} \delta_i w_i \hat{\epsilon}_i, \quad (1.1)$$

where $\hat{Y}$ is the prediction based on a regression model for $E(Y|\mathbf{X})$ which is fit using the complete cases, $w = 1/\hat{\pi}(\mathbf{X}_i)$ is the inverse of the estimated probabilities of being observed that is often computed using a regression model for $\pi(\mathbf{X})$, and $\hat{\epsilon} = Y - \hat{Y}$. On the right hand side (RHS) of (1.1), the first term is equivalent to imputing all Y values using a model for $E(Y|\mathbf{X})$, and the second term is a sum of inverse probability-weighted prediction errors due to the model for $E(Y|\mathbf{X})$ based on the weights estimated using $\pi(\mathbf{X})$. If the model for $E(Y|\mathbf{X})$ is correctly specified, then the second term converges to 0 and $\hat{\mu}_{CE}$ converges to μ. If the model for $\pi(\mathbf{X})$ is correctly specified, then one can show that the second term consistently removes any bias that may be associated with the first term and hence $\hat{\mu}_{CE}$ still converges to μ. As a result, $\hat{\mu}_{CE}$ is consistent if either of the two models is correctly specified.

Other doubly robust estimators have been introduced that use a parametric model to impute missing values and inverse probability-weighted prediction errors to correct potential bias that is associated with the parametric model for imputation. In particular, doubly robust methods were extended to regression settings (Robins, Rotnitzky, and Zhao (1994)) and repeated measurement data (Robins, Rotnitzky, and Zhao (1995)). In the context of estimating a population mean,

Qin, Shao, and Zhang (2008) and Cao, Tsiatis, and Davidian (2009) proposed two elegant approaches to improve the efficiency of the CE when the imputation model is incorrectly specified, and their methods achieve the semiparametric efficient lower bound when both models are correctly specified. During the review process, a recent related work by Hu, Follmann, and Qin (2010) was brought to our attention; they extended the CE estimator through the use of a nonparametric imputation model, where the dimension of the covariates is reduced through a parametric working index.

The double-robustness property of the CE and its extensions, though attractive, has its limitations. If one working model is misspecified, especially if it is seriously misspecified, a doubly robust estimator, although consistent, can have increased bias or variance in small samples. When both models are misspecified, a doubly robust estimator can underperform other estimators that are not doubly robust (Kang and Schafer (2007)). In addition, the inverse probability weighting step can be sensitive to missing probabilities that are close to 1. Therefore, it is desirable to develop an inference procedure, that reduces the impact of the misspecification of both working models, allows us to select and rely more heavily on the working model that is correctly specified, and is less sensitive to missing probabilities that are close to 1.

1.4. Doubly robust multiple imputation

We propose a new nonparametric MI method, and the method alleviates the curse of dimensionality, that limits the usefulness of existing nonparametric imputation methods. The proposed method is doubly robust, and differs from the CE and its extensions in that it does not use inverse-probability weighting. Our method has several advantages. First, it is more robust to two misspecified working models, being nonparametric. Second, it lessens the adverse impact of extreme missing probabilities. The method avoids the inverse probability weighting and relies only on imputation based on two working models; since one imputation creates only one pseudo observation for each observation with missing values, its impact is considerably less than that of inverse probability weighting in the presence of extreme missing probabilities. We also propose a new sensitivity analysis for empirically evaluating the validity of working models through varying weights that are used to define similarity between observations based on two working models. We note that for the CE and its related estimators the existing sensitivity analyses primarily focus on the impact of non-ignorable missingness (Rotnitzky, Robins, and Scharfstein (1998); Scharfstein, Rotnitzky, and Robins (1999)), and one that is similar to ours is neither available nor obvious. The proposed sensitivity analysis allows investigators to select optimal weights so that the resulting estimator relies completely or more heavily on the working model

that is likely to be correctly specified to achieve improved efficiency. Furthermore, the use of two weights allows investigators to incorporate their prior beliefs on the validity of two working models. In summary, our approach is intended to combine the strengths of nonparametric imputation methods and the CE method, and to overcome their respective limitations. Our main goal is not to achieve a gain in efficiency; thus, we primarily focus on comparing our approach with existing imputation methods and the CE.

The rest of the paper is organized as follows. In Section 2, we present the doubly robust nonparametric MI approach and its sensitivity analysis. In Section 3, we evaluate finite sample performance in simulation studies. In Section 4, we illustrate the proposed approach using data from a colorectal adenoma study. We conclude with some discussion in Section 5.

2. The Methodology

We first introduce the working models, then describe in detail the doubly robust nonparametric multiple imputation procedures.

2.1. Working models and predictive scores

In order to use the fully observed $\mathbf{X}$ to define an imputing set for each observation with missing Y, we consider two working models. Based on the idea of predictive mean matching (Rubin (1987)), the first model is for the outcome Y,

$$E(Y|\mathbf{X_o}) = l_1(\mathbf{X}_o, \boldsymbol{\beta}), \tag{2.1}$$

where l_1 is a specified real-valued smooth function, $\mathbf{X}_o$ is a set of p_1 observed covariates, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_1})^T$ is a vector of regression coefficients. Here l_1 is considered a predictive score of Y, for example, one can use the linear regression model, $l_1(\mathbf{X}_o, \boldsymbol{\beta}) = \boldsymbol{\beta}^T\mathbf{X}_o$. When the working model (2.1) for Y is correctly specified, an imputing set for each missing Y defined through the predictive score can lead to an improvement in efficiency when the missing mechanism is MCAR, i.e., $E(\delta|\mathbf{X}, Y) = E(\delta)$, and a reduction in bias when missing mechanism is MAR, i.e., $E(\delta|\mathbf{X}, Y) = E(\delta|\mathbf{X})$. If the working regression model for Y is misspecified, bias may remain even under a MAR mechanism. Hence, based on the idea of propensity score matching (Rosenbaum and Rubin (1983)), we take a model for predicting missingness to be

$$E(\delta|\mathbf{X}_m) = l_2(\mathbf{X}_m, \boldsymbol{\alpha}), \tag{2.2}$$

where l_2 is a specified real-valued smooth function, $\mathbf{X}_m$ is a set of p_2 observed covariates, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{p_2})^T$ is a vector of regression coefficients. Here l_2 is considered a predictive score of δ, for example, one can use the logistic

regression model, $l_2(\mathbf{X}_m, \boldsymbol{\alpha}) = \exp(\boldsymbol{\alpha}^T\mathbf{X}_m)/(1+\exp(\boldsymbol{\alpha}^T\mathbf{X}_m))$. We note that more complicated models can be used in (2.1) and (2.2). For instance, when p_1 and p_2 are large, a Lasso regression (Tibshirani (1996)) can be used for Model (2.1), and a generalized boosted model (GBM) (McCaffrey, Ridgewar and Morral (2004)) can be used for Model (2.2). The functions $l_1(\mathbf{X}_o)$ and $l_2(\mathbf{X}_m)$ may include higher order terms of $\mathbf{X}$. We denote the estimators based on (2.1) and (2.2) by $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$; throughout, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ are assumed to be $\sqrt{n}$-consistent M-estimators or Z-estimators (van der Vaart (1998)). The incorrect specification of working models can be in the functional forms of l_1 and l_2 or in the set of covariates included. In practical applications, it is difficult to correctly choose a model for $E(Y|\mathbf{X_o}) = l_1(\mathbf{X}_o, \boldsymbol{\beta})$ and there is no guarantee that the assumed model is correct. Our hope is that the proposed method can improve estimation efficiency if the assumed model is reasonably good, though not perfect.

Let $Z_1 = l_1(\mathbf{X}_o, \boldsymbol{\beta})$ and $Z_2 = l_2(\mathbf{X}_m, \boldsymbol{\alpha})$. After (2.1) and (2.2) are fit using methods such as the maximum likelihood estimation or estimation equations that achieve $\sqrt{n}$-consistency, the estimated predictive scores are $(\hat{Z}_1, \hat{Z}_2) = \{l_1(\mathbf{X}_o, \hat{\boldsymbol{\beta}}), l_2(\mathbf{X}_m, \hat{\boldsymbol{\alpha}})\}$. Alternatively, one could take the predictive scores as a monotonic transformation of (Z_1, Z_2). For example, if $l_2(\mathbf{X}_m, \boldsymbol{\alpha}) = \exp(\boldsymbol{\alpha}^T\mathbf{X}_m)$ $/(1+\exp(\boldsymbol{\alpha}^T\mathbf{X}_m))$, the linear combination $\boldsymbol{\alpha}^T\mathbf{X}_m$ could be taken as the predictive score for δ. The proposed strategy summarizes the multi-dimensional $\mathbf{X}$ with a two-dimensional predictive score, $\mathbf{Z} \equiv (Z_1, Z_2)$. In the presence of only one predictive covariate, the predictive score is the covariate itself, and there is no need to fit the two working models.

2.2. Multiple imputation (MI) estimator

To stabilize the imputation, each predictive score is standardized by subtracting its mean and dividing by its standard deviation; the resulting score is denoted by $\mathbf{S} \equiv (S_1, S_2)$. Given $\mathbf{S}$, for each subject with missing Y we create an imputing set that consists of observed responses from subjects who are similar. Specifically, $\mathbf{S}$ is used to select the imputing set by calculating the distance between subjects as

$$d(i,j) = \left\{\omega_o[S_1(i) - S_1(j)]^2 + \omega_m[S_2(i) - S_2(j)]^2\right\}^{1/2}, \tag{2.3}$$

where ω_o and ω_m are non-negative weights for the predictive scores from (2.1) and (2.2), respectively, satisfying $\omega_o + \omega_m = 1$. This choice of the weights (ω_o, ω_m) can reflect the confidence of investigators on each working model. While the double-robustness property no longer holds when ω_o or ω_m is 1, such weights are useful in sensitivity analysis, as will be illustrated in Section 2.5. For each subject i with missing Y, the distance $d(i,j)$ is used to define a set of a neighborhood,

denoted by $R_K(i)$, that consists of K subjects who have the smallest K distances (d) from subject i.

Given the imputing sets, we propose a multiple imputation (MI) estimator for the parameter of interest, μ. In the lth imputation, given the $R_K(i)$ for each subject i with missing outcome, a Y_i^* is randomly drawn with equal probability from $R_K(i)$ to replace the missing Y for subject i. We repeat this step for all subjects with missing Y, and let $\{\tilde{Y}_i(l) = \delta_i Y_i + (1-\delta_i)Y_i^*(l)\ (i = 1, \ldots, n)\}$ and $\hat{\mu}(l) = \sum_{i=1}^n \tilde{Y}_i(l)$ be the l^{th} imputed data set and the associated mean estimator, respectively. The imputation scheme is independently repeated L times to obtain L imputed data sets, and the subsequent analysis of multiple imputed data sets follows well-established rules in Rubin (1987) and Little and Rubin (2002). The final MI estimator of μ is

$$\hat{\mu}_{MI} = \frac{1}{L}\sum_{l=1}^{L}\hat{\mu}(l). \tag{2.4}$$

We refer to this method as $MI(K, \omega_o, \omega_m)$, where K is the number of the nearest neighbors, and ω_o and ω_m are the weights used to define the distance in (2.3).

2.3. Theoretical properties of MI estimator

We set forth the asymptotic properties of the proposed MI estimator as $n \to \infty$ and $L \to \infty$; a sketch of the proofs is provided in Appendix. Let $\to_p$ denote convergence in probability.

Proposition 1. *Under Conditions* (B1) *and* (B2) *in Appendix, there exist* $\boldsymbol{\beta}^0$ *and* $\boldsymbol{\alpha}^0$ *such that* $\hat{\boldsymbol{\beta}} \to_p \boldsymbol{\beta}^0$ *and* $\hat{\boldsymbol{\alpha}} \to_p \boldsymbol{\alpha}^0$.

Proposition 1 implies that the limits of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ exist even if both working models are misspecified. When the working models (2.1) and (2.2) are correctly specified, then $\boldsymbol{\beta}^0$ and $\boldsymbol{\alpha}^0$ are the true parameter values. Take the true predictive scores evaluated at the limits of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ as $\mathbf{Z}^0 \equiv (Z_1^0, Z_2^0)) = \{l_1(\mathbf{X}_o, \boldsymbol{\beta}^0), l_2(\mathbf{X}_m, \boldsymbol{\alpha}^0)\}$.

Proposition 2. *If* Y *is independent of* δ *conditional on* $\mathbf{X}$ *and if either* (2.1) *or* (2.2) *is correctly specified, then* $E(Y|\delta, \mathbf{Z}^0) = E(Y|\mathbf{Z}^0)$.

Note that the result in Proposition 2 is weaker than the conditional independence between Y and δ given $\mathbf{Z}$, as (2.1) is postulated on the mean of Y only, not on the distribution of Y.

We consider here the multiple imputation estimator computed using $\mathbf{Z}^0$ instead of $\hat{\mathbf{Z}}$, denoted by $\hat{\mu}_{MI}^0$. Take $\mu(\mathbf{Z}^0) = E(Y|\mathbf{Z}^0)$, $\pi(\mathbf{Z}^0) = Pr(\delta = 1|\mathbf{Z}^0)$, and $\sigma^2(\mathbf{Z}^0) = var(Y|\mathbf{Z}^0)$.

Theorem 1. *Under Conditions* (A1)−(A3) *in Appendix, if* ω_o *and* ω_m *are positive and either* (2.1) *or* (2.2) *is correctly specified,* $\sqrt{n}(\hat{\mu}^0_{MI}-\mu)$ *has an asymptotic normal distribution with mean* 0 *and variance*

$$\sigma^2_{MI} = \sigma^2_1 + \sigma^2_2 + \sigma^2_3 + 2\sigma_{23},$$

where

$$\sigma^2_1 = var[\mu(\mathbf{Z}^0)],\ \sigma^2_2 = E\left[var\left(\delta\{Y-\mu(\mathbf{Z}^0)\}|\mathbf{Z}^0\right)\right],$$
$$\sigma^2_3 = E\left[\frac{[1-\pi(\mathbf{Z}^0)]^2}{\pi(\mathbf{Z}^0)^2} var\left[\delta\{Y-\mu(\mathbf{Z}^0)\}|\mathbf{Z}^0\right]\right],$$

and

$$\sigma_{23} = E\left[\frac{1-\pi(\mathbf{Z}^0)}{\pi(\mathbf{Z}^0)} var\left[\delta\{Y-\mu(\mathbf{Z}^0)\}|\mathbf{Z}^0\right]\right].$$

Theorem 1 implies that $\hat{\mu}^0_{MI}$ is doubly robust and achieves a $\sqrt{n}$ convergence rate. We note that the results in Theorem 1 hold for all fixed positive weights; this is analogous to the results using kernel methods (Cheng (1994); Aerts et al. (2002)): the asymptotic results do not depend on the specific form of a kernel function. In finite samples, the impact of varying weights can be appreciable, as seen later in simulation studies. In Theorem 1, we do not need to specify the full conditional distribution of Y given $\mathbf{X}$ in (2), and Y can be continuous or discrete. Given additional conditions, one can simplify the asymptotic variance in Theorem 1.

Corollary 1. *If* ω_o *and* ω_m *are positive and either* (2.1) *or* (2.2) *is correctly specified, and if* Y *is independent of* δ *given* $\mathbf{Z}^0$ *or* $E(Y|\delta,\mathbf{Z}^0) = E(Y|\mathbf{Z}^0)$ *and* $E(Y^2|\delta,\mathbf{Z}^0) = E(Y^2|\mathbf{Z}^0)$, $\sqrt{n}(\hat{\mu}^0_{MI}-\mu)$ *is asymptotically normal with mean* 0 *and variance* $\sigma^2_{MI} = var[\mu(\mathbf{Z}^0)] + E[\sigma^2(\mathbf{Z}^0)/\pi(\mathbf{Z}^0)]$.

Two remarks are in order. First, when a monotonic transformation of $\mathbf{Z}^0$ is defined as the predictive scores, the asymptotic results in Theorem 1 still hold. In our numerical examples in Sections 3 and 4, predictive scores are defined as linear combinations of the covariates (allowing possibly higher order terms of the covariates to be included) in both the linear and logistic regression models. Second, since $\boldsymbol{\beta}^0$ and $\boldsymbol{\alpha}^0$ are unknown in practice, we need to replace them with their $\sqrt{n}$-consistent estimators ($\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$) to compute $\hat{\mathbf{Z}}$ and subsequently the MI estimator $\hat{\mu}_{MI}$. Using the influence functions for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$, and similar arguments as in the proof of Theorem 1, it is straightforward to show that $\hat{\mu}_{MI}$ has the same asymptotic normal distribution as in Theorem 1.

We can rewrite the asymptotic variance of $\hat{\mu}_{MI}$ in Corollary 1 as $n^{-1}(var(Y) + E[\{\pi(\mathbf{Z}^0)^{-1}-1\}\sigma^2(\mathbf{Z}^0)])$. When both working models are correctly

specified, the asymptotic variance of $\hat{\mu}_{MI}$ reduces to $n^{-1}(var(Y) + E[\{\pi(\mathbf{X})^{-1} - 1\}\sigma^2(\mathbf{X})])$, which is the same as the asymptotic variance of $\hat{\theta}_{CE}$ as shown in Table 1 of Hu, Follmann, and Qin (2010). Furthermore, when (2) is correctly specified and (3) is incorrectly specified, one can perform sensitivity analysis as described in Section 2.5, likely leading to a $\hat{\mu}_{MI}$ that relies only on the correctly specified working model for $E(Y|\mathbf{X})$; as shown in Tsiatis and Davidian (2007), this estimator is optimal and the propensity score is not needed.

As shown in Theorem 1 and Corollary 1, the formula for the asymptotic variance is fairly complicated and involves the density functions of the estimated predictive scores ($\hat{\mathbf{Z}}$). In practice, these density functions are often estimated using nonparametric methods, so the practical usefulness of the asymptotic variance is limited in small samples, in particular. We propose a more convenient alternative for estimating the variance of the proposed estimator, the well-established method (Rubin (1987); Little and Rubin (2002)) for estimating the variance of an MI estimator through combining within-imputation and between-imputation variances. However, it follows from Little and Rubin (2002) that the MI procedure in Section 2.2 is improper in the sense that it fails to incorporate the variability of estimating $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$. As a result, the standard method cannot be directly applied.

2.4. Bootstrap multiple imputation

To overcome the difficulty of estimating the variance of $\hat{\mu}_{MI}$, we incorporate a bootstrap step. In the lth imputation, it consists of the following steps.

1. Draw a random sample of equal size with replacement from the original data set, fit models (2.1) and (2.2) using this bootstrap sample, and compute $\mathbf{S} = (S_1^{(l)}, S_2^{(l)})$.
2. Compute the distance between a subject with a missing outcome, say subject i, and all other subjects that have an observed outcome, in the bootstrap sample as defined above. The imputing set for subject i is the nearest neighborhood $R_K^{(l)}(i)$ consisting of K subjects in the bootstrap sample with the K smallest distances from subject i. Draw a value $Y_{Bi}^*(l)$ for subject i from $R_K^{(l)}(i)$.
3. Take $\tilde{Y}_{Bi}(l) = \delta_i Y_i + (1 - \delta_i)Y_{Bi}^*(l), i = 1, \ldots, n$, and $\hat{\mu}_B(l) = \sum_{i=1}^n \tilde{Y}_{Bi}(l)$ be the l^{th} bootstrap imputed data set and the associated mean estimator, respectively.

Repeating the bootstrap imputation L times, the final bootstrap MI estimator is $\hat{\mu}_{MIB} = (1/L)\sum_{l=1}^L \hat{\mu}_B(l)$, which is referred to as $MIB(K, \omega_o, \omega_m)$. The bootstrap MI is a proper imputation (Little and Rubin (2002)) and hence its variance can be readily estimated using the sum of a between-imputation and a within-imputation component. The addition of a bootstrap step has been shown to

allow the estimation of the variance of MI estimators in other settings (Heitjan and Little (1991)); Rubin and Schenker (1991)). In our experience, $L = 5$ imputations suffice to achieve good performances in finite samples. We note that the bootstrap MI method in Aerts et al. (2002) uses a different bootstrap scheme that only resamples the complete observations, whereas our bootstrap scheme allows a resampling of the observations with missing values. In practice, non-convergence may arise when repeatedly fitting working models, say the logistic regression model, in bootstrap samples; when this happens, the bootstrap samples with the non-convergence issue are discarded.

2.5. Sensitivity analysis

The choice of (ω_o, ω_m) plays an important role in computing $MIB(K, \omega_o, \omega_m)$, and multiple estimates can be obtained using different weights. As a natural extension, a sensitivity analysis can be performed to evaluate the validity of (2.1) and (2.2). Specifically, since MIB(K, ω_o, ω_m) with nonzero weights are doubly robust and the MIB$(K, 1, 0)$ and MIB$(K, 0, 1)$ estimators are not, the differences between MIB(K, ω_o, ω_m) with nonzero weights, MIB$(K, 1, 0)$, and MIB$(K, 0, 1)$ can inform on the validity of both working models, and hence provide a justification for a sensitivity analysis. For example, if the working model (2.1) is correctly specified and (2.2) is not, then one can expect a MIB(K, ω_o, ω_m) estimator with nonzero weights to be similar to the MIB$(K, 1, 0)$ estimator, with both estimators different from the MIB$(K, 0, 1)$ estimator due to its bias. In this case, the MIB$(K, 1, 0)$ estimator may be preferred to a MIB(K, ω_o, ω_m) estimator with nonzero weights, as the use of a misspecified working model (2.2) may introduce extra noise to estimation. If the MIB estimates do not vary considerably when changing the values of weights from one extreme $(\omega_o = 1, \omega_m = 0)$ to another $(\omega_o = 0, \omega_m = 1)$, one might have more confidence in the results and choose the optimal weight $(\omega_o = 1, \omega_m = 0)$. In addition, the specification of (ω_o, ω_m) provides a natural way to incorporate prior beliefs on the validity of the two working models.

3. Simulation Studies

We conducted simulations to evaluate the finite sample performance and, in particular, the impact of incorrect specification of one or both working models and the choice of weights (ω_o, ω_m). The following estimators are compared: the sample mean of observed Y values (CC); the calibration estimator (CE) proposed by Cassel, Sarndal, and Wretman (1976); a parametric MI estimator (PMI), where a regression model with the fully observed covariates is fit and then used to draw imputes for each missing observation; the proposed bootstrap nonparametric MI estimator (MIB) $\hat{\mu}_{MIB}$. The CE, PMI, and MIB estimators

involve fitting a regression model for Y, and the CE and MIB estimators also involve fitting a regression model for δ. In our simulation studies and data example, working models were fit using the method of maximum likelihood.

Five fully observed covariates ($\mathbf{X} = (X_1, \ldots, X_5)$) were generated from independent uniform distributions on $(-1, 1)$. For the outcome of interest (Y), two true models were considered: Model (O1) where, conditional on $\mathbf{X}$, Y was generated from a normal distribution with a mean $E(Y|\mathbf{X}) = 10 + 2X_1 - 2X_2 + 3X_3 - 3X_4 + 1.5X_5$ and a variance of 9; Model (O2) where, conditional on $\mathbf{X}$, $\log(Y)$ was normal with a mean of $0.5 + 0.5X_1 - X_2 + 1.5X_3 - 2X_4 + 0.5X_5$ and a variance of 3. For the missingness indicator (δ), two true models were considered: Model (M1) where δ was generated from a logit model, $logit[Pr(\delta = 1|\mathbf{X})] = 0.5X_1 - X_2 + X_3 - X_4 + X_5$; Model (M2) where δ was generated from another logit model, $logit[Pr(\delta = 1|\mathbf{X})] = 0.5 + 2X_1 - 4X_2 + 2X_3 - 2X_4 + 2X_5$. Model (M1) generates missing probabilities that are mostly bounded away from 1, whereas Model (M2) generates more missing probabilities that are close to 1; specifically, the probability of being missing is greater than 0.95 in 0.5% of observations under Model (M1) and in 15.5% of observations under Model (M2). Simulations were conducted for combinations of the true models for Y and δ, and the following incorrect working models were used: only three predictors, X_1, X_2, and X_3, were included in the working models for Y and δ, denoted by Model (O1W), (O2W), (M1W), and (M2W), respectively; when the true model for Y is Model (O2), Model (O1) was also used as an incorrect working model (2.1). For each simulation scenario, the following measures were evaluated using 1,000 Monte Carlo data sets: the average relative bias (RB) computed using the ratio of the bias to the absolute value of the nonzero true value; the average standard error (SE) computed using a bootstrap method for CE and by combining the within and between variances for PMI and MIB; the mean squared error (MSE); the coverage rate of 95% Wald confidence intervals (CI).

The MIB estimators were computed using five different sets of values for (ω_o, ω_m): (1,0), (0.8,0.2), (0.5,0.5), (0.2,0.8) and (0,1). Note that MIB$(K, 1, 0)$ is similar to the local multiple imputation estimator (LMI) (Aerts et al. (2002)); they both use only the outcome prediction model. Still, as the number of covariates increases, LMI is subject to the curse of dimensionality, and MIB$(K, 1, 0)$ is not.

The simulation results for $n = 400$ are reported in Tables 1−3 in which $K = 3$ is used for the MIB estimators. Note that the CC estimator exhibits substantial bias in all cases. In Table 1, the true models for Y and δ are Models (O1) and (M1), respectively, and the missing probabilities are moderate. When both working models are correctly specified, the bias is negligible for all estimators including the MIB estimators with different weighting schemes; among them, the

Table 1. Simulation results when true models are Model (O1) for Y and (M1) for δ with $\mu = 10$ and $n = 400$. MIB(K,ω_o,ω_m) denotes the bootstrap MI method using K-nearest neighbors and weights ω_o and ω_m; $L = 5$ imputed datasets were used.

Method	RB(%)	SD	SE	MSE	CR(%)
CC	13.81	0.290	0.287	1.99	0.2
Correct Working Models for Both Y and δ					
CE	-0.03	0.305	0.300	0.090	94.5
PMI	0.01	0.291	0.243	0.059	89.2
MIB(3,1.0,0.0)	0.60	0.307	0.307	0.098	94.7
MIB(3,0.8,0.2)	0.60	0.317	0.305	0.097	93.3
MIB(3,0.5,0.5)	0.64	0.311	0.307	0.098	94.3
MIB(3,0.2,0.8)	0.68	0.314	0.310	0.101	93.6
MIB(3,0.0,1.0)	1.07	0.319	0.331	0.121	94.9
Wrong Working Model for Y only (O1W)					
CE	0.02	0.343	0.362	0.131	96.2
PMI	6.68	0.295	0.245	0.506	29.2
MIB(3,1.0,0.0)	6.94	0.307	0.303	0.573	44.8
MIB(3,0.8,0.2)	1.86	0.306	0.298	0.123	91.4
MIB(3,0.5,0.5)	1.39	0.311	0.303	0.111	93.0
MIB(3,0.2,0.8)	1.10	0.311	0.310	0.108	94.7
MIB(3,0.0,1.0)	1.07	0.321	0.328	0.119	94.1
Wrong Working Model for δ only (M1W)					
CE	-0.01	0.289	0.275	0.076	93.5
MIB(3,1.0,0.0)	0.60	0.305	0.307	0.098	94.5
MIB(3,0.8,0.2)	0.84	0.308	0.297	0.095	93.1
MIB(3,0.5,0.5)	1.15	0.303	0.295	0.100	94.2
MIB(3,0.2,0.8)	1.70	0.304	0.294	0.115	90.7
MIB(3,0.0,1.0)	7.06	0.313	0.309	0.594	43.3

PMI estimator has the worst coverage rate. While the bias is negligible for all MIB estimators with non-zero weights, the MIB method leads to an even smaller bias when a larger weight is assigned to the working model for Y (ω_o). When only the working model for Y is misspecified as Model (O1W), the PMI and MIB(3,1,0) estimators exhibit considerable bias, both of which rely solely on the correct specification of the working model for Y. The other four MIB estimators and CE estimator show negligible bias. In this case, as the weight increases from 0.2 to 1 for the correct working model for δ, the bias of MIB estimator decreases slightly; these MIB estimators also have slightly lower MSE compared to the CE estimator. Similarly, when only the working model for δ is misspecified as Model (M1W), the MIB(3,0,1) exhibits considerable bias due to its sole reliance on the working model for δ, whereas the other four MIB estimators and CE estimators exhibit negligible bias. In this case, the CE estimator has slightly lower MSE

Table 2. Simulation results when true models are Model (O1) for Y and Model (M2) for δ, with $\mu = 10$, and $n = 400$.

Method	RB(%)	SD	SE	MSE	CR(%)
CC	17.47	0.256	0.259	3.119	0.0
Correct Working Models for Both					
CE	-0.6	1.640	0.515	0.269	91.7
PMI	0.04	0.323	0.241	0.058	87.0
MIB(3,1.0,0.0)	1.80	0.365	0.375	0.173	90.0
MIB(3,0.8,0.2)	1.83	0.396	0.375	0.174	88.6
MIB(3,0.5,0.5)	2.04	0.420	0.396	0.198	89.3
MIB(3,0.2,0.8)	2.40	0.439	0.420	0.234	88.8
MIB(3,0.0,1.0)	2.99	0.462	0.487	0.327	90.3
Wrong Working Model for Y only (O1W)					
CE	-0.57	3.659	0.759	0.579	79.9
PMI	9.15	0.302	0.234	0.892	7.1
MIB(3,1.0,0.0)	9.92	0.314	0.327	1.091	24.5
MIB(3,0.8,0.2)	4.30	0.366	0.354	0.310	77.8
MIB(3,0.5,0.5)	3.61	0.404	0.382	0.276	82.6
MIB(3,0.2,0.8)	3.13	0.430	0.408	0.264	86.3
MIB(3,0.0,1.0)	2.97	0.463	0.488	0.328	90.6
Wrong Working Model for δ only (M2W)					
CE	0.02	0.377	0.334	0.112	92.6
MIB(3,1.0,0.0)	1.78	0.366	0.374	0.172	91.9
MIB(3,0.8,0.2)	2.30	0.381	0.355	0.179	87.8
MIB(3,0.5,0.5)	3.09	0.392	0.359	0.224	84.7
MIB(3,0.2,0.8)	4.23	0.387	0.362	0.310	78.6
MIB(3,0.0,1.0)	10.72	0.373	0.373	1.288	31.5

compared to the MIB estimators. A similar pattern regarding the impact of weights on bias is also observed in Tables 2−3, which indicates that a sensitivity analysis using different weights is useful in choosing better weighting schemes. In addition, the impact of weights on SE is minimal for MIB estimators with nonzero weights when the missing probabilities are moderate. In all cases considered in Table 1, the doubly robust MIB estimators achieve similar performance in terms of MSE when compared to the CE estimator. In addition, the doubly robust MIB estimators show slightly larger bias and MSE when the working model for Y is misspecified compared to when the working model for δ is misspecified.

In Table 2, the true models for Y and δ are Models (O1) and (M2), respectively. The true outcome model is the same as in Table 1, but the missing probabilities are more extreme. Since the estimated missing probabilities are unstable, the performance of all estimators degrades, though to different degrees. When both working models are correctly specified, or only the working model for Y is misspecified, the CE estimator has considerably larger MSE than the

MIB estimators with nonzero weights, and its SE substantially underestimates the sampling standard deviation (SD) indicating that the CE estimate is not stable. When only the working model for δ is misspecified, the weights for CE are stabilized while the correct working model for Y protects CE from being inconsistent; as a result, its performance is slightly better than that of MIB estimators. Also in this case, the impact of weights on SE is appreciable for MIB estimators with nonzero weights due to unstable estimated missing probabilities; specifically, a larger weight for the predictive score for δ tends to result in larger SD. The results regarding PMI and the impact of weighting on bias are similar to what are observed in Table 1.

Table 3 presents the simulation results when Y was generated from Model (O2), and δ was generated from Model (M1). When both working models are correctly specified, the results are similar to those in Tables 1 and 2 and hence are not included in Table 3. Two incorrect working models were used for Y: Model (O2W) which used a wrong set of covariates, and Model (O1) which used the correct set of covariates but assumed an incorrect mean function. In addition, we considered a case where both working models were misspecified. Since Y does not follow a normal distribution, we also computed the sample mean of all Y values, which is regarded as the gold standard (GS), and the coverage rate for GS is shown to be somewhat below the nominal level. When the working model for Y is misspecified as Model (O2W), CE achieves satisfactory performance; when the working model for Y is misspecified as Model (O1), CE shows appreciable bias and larger MSE compared to the MIB estimators with nonzero weights. In both cases, the MIB estimators with nonzero weights achieve satisfactory performance and PMI shows substantial bias. Interestingly, the MIB(3,1,0) estimator exhibits substantial bias when the incorrect working model (O2W) is used, whereas it shows negligible bias when the incorrect working model (O1) is used; this suggests that the MIB(3,1,0) estimator is robust to the misspecification of the working model for Y if the correct set of covariates are included. As discussed previously, the MIB method is nonparametric and only uses the predictive scores to evaluate the similarity between subjects, hence its dependence on two working models is weaker than that of the CE estimator. As long as the estimated predictive scores ($\hat{\mathbf{Z}}$) are highly correlated with the true scores, the MIB method, say, MIB(3,1,0), is expected to work. Similarly, when using the two incorrect working models (O1) and (M1W), the MIB estimators with nonzero weights still achieve performances that are comparable to GS, whereas the CE estimator shows substantial bias and coverage well below the nominal level. When the working model for δ is misspecified as Model (M1W), the CE estimator also exhibits appreciable and greater bias compared to the results in Tables 1 and 2. We note that the coverage rates of the CE estimator can be misleading in this case, since it usually exhibits

Table 3. Simulation results with true model (O2) for Y and (M1) for δ with $\mu = 8.932$, and $n = 400$.

Method	RB(%)	SD	SE	MSE	CR(%)
GS	0.87	1.560	1.418	2.017	89.4
CC	54.69	2.988	2.610	30.676	58.0
Wrong Working Model for Y only (O2W)					
CE	-0.98	1.638	1.523	2.327	90.2
PMI	24.43	2.034	2.176	9.497	94.0
MIB(3,1.0,0.0)	19.75	2.024	1.819	6.421	93.6
MIB(3,0.8,0.2)	0.94	1.686	1.527	2.339	91.9
MIB(3,0.5,0.5)	0.23	1.660	1.510	2.281	90.5
MIB(3,0.2,0.8)	-0.16	1.656	1.504	2.262	90.1
MIB(3,0.0,1.0)	0.01	1.658	1.530	2.341	90.9
Wrong Working Model for Y only (O1)					
CE	-2.99	1.947	2.102	4.490	91.3
PMI	38.61	1.605	2.267	17.034	65.1
MIB(3,1.0,0.0)	-0.20	1.661	1.495	2.235	89.6
MIB(3,0.8,0.2)	-1.14	1.629	1.472	2.177	89.7
MIB(3,0.5,0.5)	-1.22	1.648	1.475	2.187	89.9
MIB(3,0.2,0.8)	-1.60	1.633	1.469	2.178	88.6
MIB(3,0.0,1.0)	0.10	1.663	1.525	2.326	90.4
Wrong Working Model for δ only (M1W)					
CE	14.88	1.958	1.843	5.163	96.3
MIB(3,1.0,0.0)	-0.30	1.664	1.493	2.230	90.3
MIB(3,0.8,0.2)	-0.36	1.677	1.486	2.209	89.6
MIB(3,0.5,0.5)	-0.21	1.658	1.482	2.197	90.2
MIB(3,0.2,0.8)	-0.09	1.653	1.483	2.199	89.6
MIB(3,0.0,1.0)	19.92	2.027	1.849	6.584	93.8
Wrong Working Models for both (O1) and (M1W)					
CE	29.62	1.380	1.933	10.737	65.5
MIB(3,1.0,0.0)	-0.20	1.665	1.493	2.229	90.5
MIB(3,0.8,0.2)	-0.47	1.664	1.478	2.186	90.1
MIB(3,0.5,0.5)	-0.31	1.662	1.480	2.191	89.7
MIB(3,0.2,0.8)	0.02	1.670	1.480	2.190	90.1
MIB(3,0.0,1.0)	19.78	2.035	2.035	6.530	94.1

large sample variation in addition to its appreciable bias. In this setting, the impact of extreme missing probabilities was also examined; it is similar to what is observed in Tables 1 and 2.

We conducted additional simulations to investigate the impact of the number of the nearest neighbors (K) and the sample size (n). As K increases, the bias of the MIB estimator increases and its SE decreases slightly (results not shown). The MSE is comparable for different K's, though $K = 3$ in general leads to slightly lower MSEs. As the sample size increases, the performances of MIB and

CE methods improve, whereas the performances of CC and PMI methods remain unsatisfactory when the mean model for Y is misspecified.

To summarize, the proposed MIB estimators achieve similar or better performance in all settings compared with other estimators considered in our simulation studies. Our results suggest that it is more important to correctly specify the working model for Y, and larger weight for the predictive score for δ can lead to larger SD for MIB estimators. Thus, it is recommended to choose larger ω_o value (say, ≥ 0.5) in the absence of prior knowledge on the working models.

4. Data Example

We illustrate the proposed method using a colorectal adenoma data set. A colorectal polyp prevention trial was conducted at the Arizona Cancer Center, in which data were collected from 1,192 patients who underwent removal of a colorectal adenoma. Demographic information such as age, gender, body mass index (BMI), and dietary information (e.g. vitamin D), based on the Arizona Food Frequency Questionnaire (AFFQ) (Martnez et al. (1999)), were collected for all participants. The dietary intake based on the AFFQ is known to be subject to measurement error. To have a more accurate measurement, an assay based on blood/tissue samples was performed to measure the dietary intake at serum level for a subpopulation of the 1,192 participants. In particular, 598 participants were selected to have their serum vitamin D levels (Y) measured. For those participants who were not selected, their serum vitamin D levels were regarded as missing data ($\delta = 0$). We were interested in estimating the mean serum vitamin D level in the overall study population. While the selection for performing the assay was not explicitly based on demographics or disease characteristics, practical constraints in the implementation of the selection procedure may well have led to an imbalance between those who were selected and who were not. To account for a potential MAR mechanism, we applied the proposed method to estimating the overall mean serum vitamin D level.

We first constructed working models for Y and δ. Based on linear regression analyses, the serum vitamin D level was shown to be significantly associated with gender and the BMI of a patient, the number of baseline adenomas, and the vitamin D intake derived from the AFFQ. Based on logistic regression analyses for the missingness, its association with the number of baseline adenomas achieves statistical significance with an estimated odds ratio (OR) 1.18 and a 95% CI (1.04, 1.34), and its association with the gender of a patient achieves marginal statistical significance with an estimated OR 1.27 and a 95% CI (0.99, 1.61). Consequently, to compute the CE and MI estimators, we included the gender and BMI of a patient, the number of baseline adenomas, and the vitamin D intake from the AFFQ as covariates to fit a linear regression model for predicting

Table 4. Estimation of the overall mean level of serum vitamin D for a colon cancer study.

Method	Estimate	SE	95% CI
CC	26.262	0.385	(25.508, 27.016)
CE	26.267	0.364	(25.554, 26.981)
MIB(3,1.0, 0.0)	26.133	0.315	(25.516, 26.751)
MIB(3,0.8, 0.2)	26.364	0.309	(25.759, 26.969)
MIB(3,0.5, 0.5)	26.249	0.500	(25.269, 27.229)
MIB(3,0.2, 0.8)	26.558	0.330	(25.911, 27.206)
MIB(3,0.0, 1.0)	26.438	1.642	(23.220, 29.656)

serum vitamin D level, and included the patient's gender and the number of baseline adenomas as covariates to fit a logistic regression model for predicting the missing probability. To compute MIB estimators with different weighting schemes, we chose $K = 3$ and $L = 5$.

The results are reported in Table 4 for CC, CE, and MIB. All methods produce a similar point estimate of the mean serum vitamin D level. The CE method produces a lower estimate (5.4% lower) of standard error (SE) compared to the CC analysis. The MIB method produces a lower SE (19.8% lower) compared to the CC analysis when a small weight (e.g. 0.2) is used for the predictive score for the missing probability. When building the working model for δ, one sees that the association between missingness and other covariates is in general weak, i.e., ORs close to 1. Thus, it is likely that the missing data mechanism is close to MCAR in this dataset. In addition, our results seem to indicate that the working model for Y is approximately correct. Consequently, a larger weight for the predictive score of the missing probability may introduce extra noise to the estimation in a single data set, which can manifest itself in the form of higher SE, larger bias, or both. In summary, the working model for the missing probability is likely incorrect, whereas the working model for the outcome is likely close to the truth. As a result, either MIB$(3, 0.8, 0.2)$ or MIB$(3, 1, 0)$ could be chosen as the estimate of the overall mean serum vitamin D level.

5. Discussion

Under MAR, we have investigated a nonparametric multiple imputation approach to estimating the marginal mean of a continuous or discrete random variable that is missing for some subjects. Working models are used to achieve two main goals: dimension reduction and double-robustness. Compared to CE and its related estimators, our approach has weak reliance on both working models in the sense that it only uses the estimated predictive scores to evaluate the similarity between subjects; along the lines of Hu, Follmann, and Qin (2010) (in particular, DEFINITION 1, p. 306), as long as our predictive scores Z_1 and Z_2

are an atom of $E(Y|\mathbf{X})$ and $E(\delta|\mathbf{X})$, respectively, the results in Section 2 still hold. Our approach also incorporates a bootstrap step, which provides a convenient way to estimate the variance of the estimator. In addition, our proposed sensitivity analysis allows investigators to incorporate prior beliefs on the validity of the working models, and, more importantly, evaluate the validity of the working models, which in turn enables investigators to choose an optimal estimator. For CE and its related estimators, a similar sensitivity analysis is lacking and it is not obvious how to develop such a sensitivity analysis. The proposed approach can be extended to other settings such as regression analysis in the presence of missing data. Based on our numerical results, we recommend that investigators always perform the sensitivity analysis and set ω_o to a larger value in the absence of strong prior knowledge.

In the context of surveys, Haziza and Beaumont (2007) proposed imputation methods based on two scores that are similar to our $\mathbf{Z}$. They proposed to use a classification algorithm to partition the sample into disjoint classes and then to impute the missing values within each class; this differs from our approach in that our method allows the K-nearest neighbors to overlap. Their approach may encounter difficulty when no obvious clusters exist in the data, and its theoretical properties are unknown. We have used a K-nearest neighbor method to allow for adaptation to the local density of the data and missing probabilities. It is a future interest to study principled approaches for selecting K as well as additional adaptations.

Acknowledgement

We would like to thank Editor Kung-Yee Liang, an associate editor, and two referees for helpful comments that greatly improved an earlier draft of this manuscript. This work was supported in part by an NIH/NCI R03 grant R03 CA130089-01A1.

Appendix: Proofs

Let $h_1(\mathbf{Z}^0)$ be the density function of $\mathbf{Z}^0$. The following regularity conditions are stated for Theorem 1 and Corollary 1.

(A1) Y has finite first and second moments, and σ_1^2, σ_2^2, σ_3^2, and σ_{23} as defined in Theorem 1 are finite.

(A2) $K/n \to 0$ and $K/\log(n) \to \infty$.

(A3) $h_1(\mathbf{Z}^0)$ and $\pi(\mathbf{Z}^0)$ are continuous and bounded away from 0 in the compact support of $\mathbf{Z}^0$.

A.1. Proof of Proposition 1

We prove Proposition 1 for $\hat{\boldsymbol{\beta}}$ only, since the arguments for $\hat{\boldsymbol{\alpha}}$ are similar. The following conditions are assumed to hold for $\hat{\boldsymbol{\beta}}$.

(B1) $\hat{\boldsymbol{\beta}}$ is the maximizer of a strictly concave objective function, $\ell_n(\boldsymbol{\beta})$, or the unique solution to a set of estimation equations, $U_n(\boldsymbol{\beta}) = 0$.

(B2) $\ell_n(\boldsymbol{\beta})$ (or $U_n(\boldsymbol{\beta})$) converges almost surely to $\ell(\boldsymbol{\beta}) = E\{\ell_n(\boldsymbol{\beta})\}$ (or $U(\boldsymbol{\beta}) = E\{U_n(\boldsymbol{\beta})\}$), uniformly in $\boldsymbol{\beta}$; $\ell(\boldsymbol{\beta})$ is strictly concave with a unique maximizer $\boldsymbol{\beta}^0$ (or $U(\boldsymbol{\beta})$ has a unique solution $\boldsymbol{\beta}^0$).

Note that Conditions (B1) and (B2) are satisfied for most regression models including linear and generalized linear models and for many estimation equations. In either case, it follows from arguments similar to those in Section 5.2 of van der Vaart (1998) that $\hat{\boldsymbol{\beta}} \to_p \boldsymbol{\beta}^0$. As discussed in Section 5.2 of van der Vaart (1998), Conditions (B1) and (B2) can be relaxed and Proposition 1 holds for most, if not all, potential working models for Y.

A.2. Proof of Proposition 2

If (2.2) is correctly specified, $Z_2^0 = l_2(\mathbf{X}_m, \boldsymbol{\alpha}^0)$ is the propensity score defined in Rosenbaum and Rubin (1983) and Proposition 2 follows from arguments similar to theirs. If (2.1) is correctly specified, then $Z_1^0 = l_1(\mathbf{X}_o, \boldsymbol{\beta}^0) = E(Y|\mathbf{X}_o)$ and

$$\begin{aligned} E(Y|\delta, \mathbf{Z}^0) &= E\left\{E(Y|\delta, \mathbf{Z}^0, \mathbf{X})|\delta, \mathbf{Z}^0\right\} = E\left\{E(Y|\mathbf{Z}^0, \mathbf{X})|\delta, \mathbf{Z}^0\right\} \\ &= E\left(Z_1^0|\delta, \mathbf{Z}^0\right) = Z_1^0 = E\left(Y|\mathbf{Z}^0\right), \end{aligned}$$

where the second equality is due to MAR. The proof is complete.

A.3. Proof of Theorem 1

Under MAR, if the weights (ω_o, ω_m) are positive and either of the working models (2.1) and (2.2) is correctly specified, then it follows from Proposition 2 that $E(Y|\mathbf{Z}^0, \delta) = E(Y|\mathbf{Z}^0)$; this implies that we can use $E(Y|\mathbf{Z}^0, \delta = 1)$ based on observed data to impute $E(Y|\mathbf{Z}^0, \delta = 0)$ for observations with missing Y. This result is used implicitly throughout the proof.

To derive the asymptotic distribution of $\hat{\mu}_{MI}^0$ with positive weights, we first consider another estimator, the K-nearest-neighbor plug-in estimator

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\Big[\delta_i Y_i + (1-\delta_i)\sum_{j\in R_K(i)}\frac{1}{K}Y_j\Big],$$

where $R_K(i)$ is the set of K nearest observed neighbors of Y_i defined using the distance in (2.3). We note that the K nearest neighbors are chosen from the subjects with observed outcomes, i.e., $\delta = 1$.

A.3.1. Asymptotic Distribution of $\sqrt{n}(\hat{\mu} - \mu)$

Consider a more general case where

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\Big[\delta_i Y_i + (1-\delta_i)\sum_{j=1}^{n}\frac{W_{ij}\delta_j}{\sum_k W_{ik}\delta_k}Y_j\Big]$$

and $W_{ij} = W(\mathbf{Z}_i^0, \mathbf{Z}_j^0)$ are consistent probability weights as defined in Stone (1977). Then we write

$$\hat{\mu} - \mu = T_1 + T_2 + T_3 + T_4,$$

where $T_1 = (1/n)\sum_{i=1}^{n}\big[\mu(\mathbf{Z}_i^0) - \mu\big]$, $T_2 = (1/n)\sum_{i=1}^{n}\delta_i\big[Y_i - \mu(\mathbf{Z}_i^0)\big]$, $T_3 = (1/n)\sum_{i=1}^{n}(1-\delta_i)\sum_{j=1}^{n}(W_{ij}\delta_j/\sum_k W_{ik}\delta_k)[Y_j - \mu(\mathbf{Z}_j^0)]$, and $T_4 = (1/n)\sum_{i=1}^{n}(1-\delta_i)\sum_{j=1}^{n}(W_{ij}\delta_j/\sum_k W_{ik}\delta_k)[\mu(\mathbf{Z}_j^0) - \mu(\mathbf{Z}_i^0)]$.

It is straightforward to show that $\sqrt{n}T_1 \to_d N(0, \sigma_1^2)$ and $\sqrt{n}T_2 \to_d N(0, \sigma_2^2)$, where $\sigma_1^2 = var[\mu(\mathbf{Z}^0)])$ and $\sigma_2^2 = E\left[var\left(\delta\{Y - \mu(\mathbf{Z}^0)\}|\mathbf{Z}^0\right)\right]$. Let

$$T_3^* = \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\sum_{j=1}^{n}\frac{1}{n}\frac{W_{ij}\delta_j}{h(\mathbf{Z}_i^0)}[Y_j - \mu(\mathbf{Z}_j^0)],$$

$$T_4^* = \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\sum_{j=1}^{n}\frac{1}{n}\frac{W_{ij}\delta_j}{h(\mathbf{Z}_i^0)}[\mu(\mathbf{Z}_j^0) - \mu(\mathbf{Z}_i^0)],$$

where $h(\mathbf{Z}_i^0) = \pi(\mathbf{Z}_i^0)h_1(\mathbf{Z}_i^0)$; $h(\mathbf{Z}_i^0)$ can be estimated by $\hat{h}(\mathbf{Z}_i^0) = \sum_k W_{ik}\delta_k/n$. Next, we show that $\sqrt{n}T_3 - \sqrt{n}T_3^* \to_p 0$ and $\sqrt{n}T_4 - \sqrt{n}T_4^* \to_p 0$. Since proofs are similar, we only focus on T_3. Appealing to previous work on the uniform convergency of nearest neighbor density estimates (Devroye and Wagner (1977)) and kernel density estimates (Silverman (1978)), it is straightforward to show the strong uniform convergency of $\hat{h}(\mathbf{Z}^0)$ to $h(\mathbf{Z}^0)$ under Conditions (A2) and (A3). Note that

$$E(T_3 - T_3^*)^2 = E\Big[\frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\sum_{j=1}^{n}\frac{1}{n}\frac{W_{ij}\delta_j}{h(\mathbf{Z}_i^0)}\left\{Y_j - \mu(\mathbf{Z}_j^0)\right\}\left\{\frac{h(\mathbf{Z}_i^0)}{\hat{h}(\mathbf{Z}_i^0)} - 1\right\}\Big]^2.$$

Following the proof of the asymptotic distribution of T_3^* that is discussed later, it can be shown that $E(\sqrt{n}T_3^*)^2 = \sigma_3^2 + o(1)$, the RHS of which is bounded. Then, given the uniform integrability of $(\sqrt{n}T_3^*)^2$ and the uniform convergence

of $\hat{h}(\mathbf{Z}^0)$, one can establish the asymptotic mean square equivalence of $\sqrt{n}T_3$ and $\sqrt{n}T_3^*$. It follows that $\sqrt{n}T_3 - \sqrt{n}T_3^* \to_p 0$. Similarly one can prove that $\sqrt{n}T_4 - \sqrt{n}T_4^* \to_p 0$.

We now show that $\sqrt{n}T_3^* \to_d N(0, \sigma_3^2)$. Take $R_i = \delta_i[Y_i - \mu(\mathbf{Z}_i^0)]$, we can reexpress T_3^* as

$$\begin{aligned} T_3^* &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(1-\delta_i)\frac{W_{ij}}{h(\mathbf{Z}_i^0)}R_j \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}\left[(1-\delta_i)\frac{W_{ij}}{h(\mathbf{Z}_i^0)}R_j + (1-\delta_j)\frac{W_{ji}}{h(\mathbf{Z}_j^0)}R_i\right] \\ &= \frac{1}{n^2}\sum_{i\neq j}H(Z_i^*, Z_j^*), \end{aligned}$$

where $Z_i^* = (Y_i, \mathbf{Z}_i^0, \delta_i)$, $H(Z_i^*, Z_j^*) = (1/2)[(1-\delta_i)(W_{ij}/h(\mathbf{Z}_i^0))R_j + (1-\delta_j)(W_{ji}/h(\mathbf{Z}_j^0))R_i]$. Now, $U = [n(n-1)]^{-1}\sum_{i\neq j}H(Z_i^*, Z_j^*)$ is a standard U-Statistic and $T_3^* = [(n-1)/n]U$. It is straightforward to show that $\sqrt{n}T_3^* - \sqrt{n}U \to_p 0$. Applying standard U-statistic theory, let the projection of U be $\hat{U}$ and then $\hat{U} = (2/n)\sum_i H_1(Z_i^*)$ where

$$\begin{aligned} H_1(Z_i^*) &= E\left[H(Z_i^*, Z_j^*)|Z_i^*\right] = \frac{1}{2}R_i E\left[\frac{W(\mathbf{Z}^0, \mathbf{Z}_i^0)}{h(\mathbf{Z}^0)}(1-\pi(\mathbf{Z}^0))|\mathbf{Z}_i^0\right] \\ &= \frac{1}{2}R_i\frac{1-\pi(\mathbf{Z}_i^0)}{\pi(\mathbf{Z}_i^0)} + O(\frac{K}{n}). \end{aligned}$$

It can be readily shown that $\mathrm{Var}\,[H_1(Z^*)] = \sigma_3^2/4 + O(K/n)$ with $\sigma_3^2 = E[\{[1-\pi(\mathbf{Z}^0)]^2/\pi(\mathbf{Z}^0)^2\}var[\delta\{Y-\mu(\mathbf{Z}^0)\}|\mathbf{Z}^0]]$. Since the $H_1(Z_i^*)$'s are mutually independent, it follows that $\sqrt{n}\hat{U} \to_d N(0, \sigma_3^2)$ and hence $\sqrt{n}T_3 \to_d N(0, \sigma_3^2)$. Similarly, one can show that $\sqrt{n}T_4 \to_p 0$.

Finally, it is straightforward to show that $T_1 \perp T_2$, $T_1 \perp \hat{U}$, and $Cov(\sqrt{n}T_2, \sqrt{n}\hat{U}) \to \sigma_{23}$, where $\sigma_{23} = E\left[[(1-\pi(\mathbf{Z}^0))/\pi(\mathbf{Z}^0)]var\left[\delta\{Y-\mu(\mathbf{Z}^0)\}|\mathbf{Z}^0\right]\right]$. It then follows that $\sqrt{n}(\hat{\mu}-\mu) \to_d N\left(0, \sigma^2\right)$, where $\sigma^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + 2\sigma_{23}$.

A.3.2. Asymptotic Distribution of $\sqrt{n}(\hat{\mu}_{MI}^0 - \mu)$

After rearranging terms, we have

$$\hat{\mu}_{MI}^0 = \frac{1}{n}\sum_{i=1}^{n}\Big[\delta_i Y_i + (1-\delta_i)\sum_{j\in R_k(i)}\frac{l_{ij}}{L}Y_j\Big],$$

where l_{ij} is the number of imputed data sets in which Y_j is used to impute Y_i,

and $\sum_j l_{ij} = L$. Then we have

$$\hat{\mu}^0_{MI} - \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\sum_{j\in R_K(i)}\left(\frac{l_{ij}}{L} - \frac{1}{K}\right)Y_j,$$

where $E(l_{ij}/L|K) = 1/K$, $\mathrm{Var}(l_{ij}/L) = (K-1)/(L\times K^2)$, and $\mathrm{Cov}(l_{ij}/L, l_{ik}/L|K) = (L\times K^2)^{-1}$. One can then show that $E[\sqrt{n}(\hat{\mu}^0_{MI} - \hat{\mu})]^2 \le (L)^{-1}C$, where $C = EY^2$ and is finite under Condition (A1). If $1/L \to 0$, then $\sqrt{n}(\hat{\mu}^0_{MI} - \hat{\mu})$ converges to 0 in L_2, and hence in probability. The asymptotic distribution of $\sqrt{n}(\hat{\mu}^0_{MI} - \mu)$ is therefore $N(0, \sigma^2_{MI})$ where $\sigma^2_{MI} = \sigma^2_1 + \sigma^2_2 + \sigma^2_3 + 2\sigma_{23}$. The proof of Theorem 1 is now complete.

Two remarks are in order. First, $\mathbf{X}$ may include both categorical or continuous variables; when $\mathbf{X}$ are all categorical in one or both working models, one or both components of $\mathbf{Z}^0$ are discrete, and continuity and compactness are then defined given the usual topology for a discrete space. It is well known that a compact discrete space is finite; as a result, it is straightforward to show that the results of Theorem 1 still hold. Secondly, the main result in Theorem 1 is proved under Condition (A3); this can be relaxed using a trimming technique similar to that in Härdle, Janssen, and Serfling (1988).

A.4. Proof of Corollary 1

If Y is independent of δ given $\mathbf{Z}^0$, or $E(Y|\delta, \mathbf{Z}^0) = E(Y|\mathbf{Z}^0)$ and $E(Y^2|\delta, \mathbf{Z}^0) = E(Y^2|\mathbf{Z}^0)$, then it is straightforward to show that $\sigma^2_2 = E[E(\delta|\mathbf{Z}^0)var\{Y - \mu(\mathbf{Z}^0)|\mathbf{Z}^0\}] = E[\pi(\mathbf{Z}^0)\sigma^2(\mathbf{Z}^0)]$. Similarly, one can show that $\sigma^2_3 = E[\sigma^2(\mathbf{Z}^0)[1 - \pi(\mathbf{Z}^0)]^2/\pi(\mathbf{Z}^0)]$ and $\sigma_{23} = E\left[\left\{1-\pi(\mathbf{Z}^0)\right\}\sigma^2(\mathbf{Z}^0)\right]$. It follows that $\sigma^2_{MI} = var[\mu(\mathbf{Z}^0)] + E[\sigma^2(\mathbf{Z}^0)/\pi(\mathbf{Z}^0)]$. The proof is complete.

References

Aerts, M., Claeskens, G., Hens, N. and Molenberghs, G. (2002). Local multiple imputation. *Biometrika* **89**, 375-388.

Cao, W., Tsiatis, A. and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* **96**, 723-734.

Cassel, C. M., Sarndal, C. E. and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* **63**, 615-620.

Cheng, P. E. (1994). Kernel estimation of distribution functions and quantities with missing data. *J. Amer. Statist. Assoc.* **89**, 81-87.

Devroye, L. P. and Wagner, T. J. (1977). The strong uniform consistency of nearest neighbor density estimates. *Ann. Statist.* **5**, 536-540.

Härdle, W., Janssen, P. and Serfling, R. (1988). Strong uniform consistency rates for estimators of conditional functionals. *Ann. Statist.* **16**, 1428-1449.

Haziza, D. and Beaumont, J. (2007). On the construction of imputation classes in surveys. *Internat. Statist. Rev.* **75**, 25-43.

Heitjan, D. F. and Little, R. J. A. (1991). Multiple imputation for the fatal accident reporting system. *Appl. Statist.* **40**, 13-29.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47**, 663-685.

Hu, Z., Follmann, D. and Qin, J. (2010). Semiparametric dimension reduction estimation for mean response with missing data. *Biometrika* **97**, 305-319.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22**, 523-539.

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

Martnez, M. E., Marshall, J. R., Graver, E., Whitacre, R., Woolf, K., Ritenbaugh, C., S. and Alberts, D. S. (1999). Reliability and validity of a self-administered food frequency questionnaire in a chemoprevention trial of adenoma recurrence. *Cancer Epidemiology, Biomarkers & Prevention* **8**, 941-946.

Matloff, N. M. (1981). Use of regression functions for improved estimation of means. *Biometrika* **68**, 685-689.

McCaffrey, D., Ridgewar, G. and Morral, A. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol. Methods* **9**, 403-425.

Qin, J., Shao, J. and Zhang, B. (2008). Efficient and doubly robust imputation for covariate-dependent missing responses. *J. Amer. Statist. Assoc.* **103**, 797-809.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89**, 846-86.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90**, 106-121.

Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41-55.

Rotnitzky, A., Robins, J. and Scharfstein, D. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J. Amer. Statist. Assoc.* **93**, 1321-1339.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Rubin, D. B. and Schenker, N. (1991). Multiple imputation in health-care databases: An overview and some applications. *Statist. Medicine* **10**, 585-598.

Scharfstein, D., Rotnitzky, A. and Robins, J. (1999). Adjusting for nonignorable dropout using semiparametric nonresponse models. *J. Amer. Statist. Assoc.* **94**, 1096-1120.

Silverman, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Statist.* **6**, 177-184.

Stone, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5**, 595-645.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Titterington, D. M. and Sedransk, J. (1989). Imputation of missing values using density estimation. *Statist. Probab. Lett.* **8**, 411-418.

Tsiatis, A. Davidian, M. (2007). Comment on 'Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data.' *Statist. Sci.* **22**, 569-573.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, New York.

Wang, Q., Linton, O. and Härdle, W. (2004). Semiparametric regression analysis with missing response at random. *J. Amer. Statist. Assoc.* **99**, 334-345.

Department of Biostatistics and Bioinformatics, Emory University, 1518 Clifton Rd. NE, Atlanta, GA, 30322, USA.

E-mail: qlong@emory.edu

Division of Epidemiology and Biostatistics, College of Public Health, University of Arizona, Tucson, AZ 85724, USA.

E-mail: phsu@azcc.arizona.edu

Department of Biostatistics, Division of Quantitative Sciences, The University of Texas MD Anderson Cancer Center, 1400 Pressler St, Unit 1411, Houston, TX 77030, USA.

E-mail: ysli@mdanderson.org

(Received March 2010; accepted November 2010)