*Journal of Survey Statistics and Methodology (2020) **0**, 1–24*

NONPARAMETRIC MASS IMPUTATION FOR DATA INTEGRATION

SIXIA CHEN*
SHU YANG
JAE KWANG KIM

Data integration combining a probability sample with another nonprobability sample is an emerging area of research in survey sampling. We consider the case when the study variable of interest is measured only in the nonprobability sample, but comparable auxiliary information is available for both data sources. We consider mass imputation for the probability sample using the nonprobability data as the training set for imputation. The parametric mass imputation is sensitive to parametric model assumptions. To develop improved and robust methods, we consider nonparametric mass imputation for data integration. In particular, we consider kernel smoothing for a low-dimensional covariate and generalized additive models for a relatively high-dimensional covariate for imputation. Asymptotic theories and variance estimation are developed. Simulation studies and real applications show the benefits of our proposed methods over parametric counterparts.

Sixia Chen is an Assistant Professor with the Department of Biostatistics and Epidemiology, The University of Oklahoma Health Sciences Center, Oklahoma City, OK 73126-0901, USA. Shu Yang is an Assistant Professor with the Department of Statistics, North Carolina State University, Raleigh, NC 27607, USA. Jae Kwang Kim is Professor with the Department of Statistics, Iowa State University, Ames, IA 50011-1090, USA

Dr S.C. is partly supported by the Oklahoma Shared Clinical and Translational Resources (U54GM104938) with an Institutional Development Award (IDeA) from National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the social views of the National Institutes of Health. Dr S.Y. is partly supported by ORAU, NSF DMS 1811245, and NCI P01 CA142538. Dr J.-K.K. is partly supported by NSF MMS-1733572.

*Address correspondence to Sixia Chen, Department of Biostatistics and Epidemiology, The University of Oklahoma Health Sciences Center, Oklahoma City, OK 73126-0901, USA; E-mail: sixia-chen@ouhsc.edu.

doi: 10.1093/jssam/smaa036 Advance access publication 0 Month 0000

KEYWORDS: Approximate Bayesian; Generalized additive model; Hybrid bootstrap; Kernel smoothing; Missingness at random; Nonprobability sample.

1. INTRODUCTION

Probability sampling is a scientific way of obtaining a representative sample from the target population. Official statistics are mostly computed based on probability samples. However, obtaining a probability sample is expensive and time-consuming, and is often subject to nonresponse. However, nonprobability samples become increasingly available and can be used to complement gold-standard probability sampling even though the scientific justification for using nonprobability samples is still limited (Keiding and Louis 2016). Because of the lack of information on sampling mechanisms, the nonprobability sample is often not representative of the target population. Thus, statistical inference from nonprobability samples without further adjustment may lead to biased results and misleading interpretations.

Data integration is one way to leverage the information from the nonprobability sample and to learn the outcome–covariate relationship. This process combines information from a probability sample with information from the nonprobability sample to obtain a valid inference for the target population (Lohr and Raghunathan 2017). We consider the case when the study variable of interest is measured only in the nonprobability sample, but comparable auxiliary information is available for both data sources. Thus, the probability sample is used to obtain the representativeness of the sample, but the measurement is made only in the nonprobability sample.

There are two main approaches for combining the probability and nonprobability samples. One approach is to use propensity weighting to improve the representativeness of the nonprobability sample (Chen, Li, and Wu 2019; Yang, Kim, and Song 2020). The other approach is to use mass imputation, which creates synthetic imputed values of the study variable for the probability sample using the nonprobability sample as a training sample for developing the imputation model. Rivers (2007) proposed using the value of the nearest neighbor for mass imputation, but did not discuss its properties theoretically. Yang and Kim (2018) filled the theoretical gap by establishing the asymptotic distribution of the nearest neighbor mass imputation estimators; however, nearest neighbor imputation estimators suffer from the curse of dimensionality. Kim, Park, Chen, and Wu (2018) proposed using regression models for mass imputation and discussed its statistical properties, including consistent variance estimation. However, such a parametric mass imputation method is subject to model misspecification bias.

In this paper, we develop nonparametric mass imputation estimators from the probability sample where the nonparametric imputation model is trained

based on the nonprobability sample. Instead of restrictive parametric modeling assumptions, we only require mild smoothness conditions for the outcome model and therefore the imputation estimators gains robustness over parametric counterparts. Moreover, the computation for nonparametric mass imputation can be implemented using off-the-shelf software packages. Although the nonparametric imputation approaches for traditional missing data problems have been considered in the literature, for example, Cheng (1994), statistical inference after nonparametric mass imputation has not been investigated. To fill in this important research gap, we establish theoretical properties of the proposed nonparametric mass imputation estimators and develop consistent variance estimators. Our framework covers Kernel regression and generalized additive models (GAMs) for imputation, but its extension to other nonparametric methods can be developed similarly.

The paper is organized as follows. In Section 2, the basic setup is introduced. Section 3 covers the proposed methods by using the Kernel regression technique. Proposed methods using generalized additive modeling and variance estimation using hybrid bootstrap and the approximate Bayesian method are presented in Section 4. In Section 5, results from two limited simulation studies are presented. In Section 6, we present a real-data application of the proposed method to analyze a nonprobability survey sample by using National Health Insurance Sharing Service (NHISS) and Korea National Health and Nutrition Examination Survey (KNHANES) data. Some discussion is presented in Section 7. All technical details are contained in the Appendix.

2. BASIC SETUP

Suppose that we have two independent samples selected from the same target population, denoted by sample A and sample B, where sample A is obtained from probability sampling and sample B is a nonprobability sample, such as a voluntary sample or a self-selected sample. In the nonprobability sample, the sample inclusion probabilities are unknown, and therefore valid analysis of the nonprobability sample is extremely difficult. To reflect the situation where the most up-to-date information is obtained from the nonprobability sample, we assume that the study variable of interest Y is observed only in the nonprobability sample B. The vector of auxiliary variable X is observed in both samples. Table 1 presents the data structure of our setup.

Because the study variable is not measured for sample A, a natural approach is to obtain the predicted values of the study variable based on the observed auxiliary information in sample A. The prediction model is trained using the full sample observations in sample B. Once the predicted values are created for all elements in sample A, we can treat these predicted values as if they are real observations and apply the Horvitz–Thompson estimation (Horvitz and Thompson 1952) using the sampling weights, nonresponse adjusted weights,

Table 1. Data Structure from Two Samples

Sample	X	Y	Type
A	✓		Probability
B	✓	✓	Nonprobability

calibration weights, or other types of adjusted weights in sample A. This method is called mass imputation for data integration. Mass imputation has been developed in the context of two-phase sampling (Breidt, McVey, and Fuller 1996; Kim and Rao 2012), but it is not fully investigated in the context of survey integration for combining the probability and nonprobability samples.

To formally describe the setup, suppose a finite population $\mathcal{F}_N = \{(x_i, y_i), i = 1, \ldots, N\}$ follows the super-population model:

$$y_i = m(x_i) + \epsilon_i, \tag{1}$$

where y_i is a scalar study variable, x_i is a d_x-dimensional covariate, $m(x_i)$ is a unknown function of x_i, and the ϵ_i's are independent errors that satisfy $E(\epsilon_i|x_i) = 0$ and $E(\epsilon_i^2|x_i) = v(x_i)$ with $v(x_i)$ as a unknown function of x_i, for $i = 1, \ldots, N$. For simplicity, we focus on estimating the population mean of y, $\theta_N = N^{-1}\sum_{i=1}^N y_i$, although our framework is applicable to other parameters such as domain means; see the simulation study in Section 5.

Given the finite population, suppose a probability sample A is selected by a sampling design. Let I_i be the sampling indicator for unit i; that is, $I_i = 1$ if unit i is selected and 0 otherwise. The corresponding first- and second-order inclusion probabilities are defined as $\pi_i = E(I_i|\mathcal{F}_N)$ and $\pi_{ij} = E(I_iI_j|\mathcal{F}_N)$ for $i \neq j$, where $E(\cdot|\mathcal{F}_N)$ is the expectation taken with respect to the sampling distribution given the finite population. Then, the design weight is $w_i = \pi_i^{-1}$. In addition, the nonprobability sample B is obtained from $U = \{1, 2, \ldots, N\}$ with the sampling indicator δ_i; that is, $\delta_i = 1$ if unit i is in sample B and 0 otherwise. We assume that the indicators I_i and δ_i are independent with each other. In contrast to sample A, where the sampling mechanism is known, the selection probability into sample B,

$$\pi_B(x, y) = \Pr(\delta = 1|x, y), \tag{2}$$

is unknown, where $\pi_B(x, y)$ denotes an unknown function of x and y. We further assume a noninformative sampling mechanism in the sense that $\pi_B(x, y) = \pi_B(x)$. The noninformativeness assumption is a strong assumption, and there is no way to check this assumption. It is not verifiable based on the observed data. See our discussion in Section 7.

When model (1) is a linear regression model $m(x) = x^{\mathrm{T}}\beta$ for some β, Kim et al. (2018) considered a mass imputation estimator

$$\widehat{\theta}_{\mathrm{MIE}} = \frac{1}{N}\sum_{i\in A} w_i \widehat{m}(x_i),$$

where $\widehat{m}(x_i) = x_i^{\mathrm{T}}\widehat{\beta}$, $\widehat{\beta} = (\sum_{i\in B} x_i x_i^{\mathrm{T}})^{-1}\sum_{i\in B} x_i y_i$, and they proposed a consistent variance estimator of $\widehat{\theta}_{\mathrm{MIE}}$ under model (1) and the noninformative sampling mechanism assumption for sample B. However, their proposed method relies on a correctly specified model for $m(x)$. In Section 3, we relax this strong assumption and propose nonparametric and semiparametric mass imputation using kernel smoothing and generalized additive modeling approaches. Without particular specification, E and V denote the expectation and variance under the randomness due to an imputation model, the marginal distribution of x, and the random sampling processes of sample A and sample B, respectively.

3. KERNEL SMOOTHING

Under noninformative sampling, $E(y_i|x_i, \delta_i = 1) = E(y_i|x_i)$; that is, the mean function is *transportable*. Motivated by Cheng (1994), we propose the following mass-imputed estimator:

$$\widehat{\theta}_{\mathrm{MIE}} = \frac{1}{N}\sum_{i\in A} w_i \widehat{m}(x_i), \tag{3}$$

where

$$\widehat{m}(x_i) = \frac{\sum_{j\in B} K_h(x_i, x_j) y_j}{\sum_{j\in B} K_h(x_i, x_j)}, \tag{4}$$

with x_i as a d_x-dimensional covariate, $K_h(x_i, x_j) = K\{h^{-1}(x_i - x_j)\}$ with K as a multivariate kernel density function including uniform, standard normal, or triangular density functions as special cases (Epanechnikov 1969) and h as the bandwidth. Wang, Graubard, Katki, and Li (2020) considered using kernel functions to improve external validity of epidemiologic cohort analyses through weighting. To discuss asymptotic properties of $\widehat{\theta}_{\mathrm{MIE}}$ in (3), we consider a sequence of finite populations and samples as described in Fuller (2011). With regularity conditions defined in Appendix A.1, we have the following theorem. The sketched proof of Theorem 1 can be found in Appendix A.2.

Theorem 1. Under the regularity conditions specified in Appendix A.1, the kernel-base mass imputation estimator in (3) satisfies

$$\widehat{\theta}_{MIE} - \tilde{\theta}_{MIE} = o_p(n^{-1/2}), \tag{5}$$

where

$$\tilde{\theta}_{MIE} = \frac{1}{N}\sum_{i\in A} w_i m(x_i) + \frac{1}{N}\sum_{i\in B} g_B(x_i)\{y_i - m(x_i)\} \tag{6}$$

with

$$g_B(x_i) = \sum_{j=1}^{N}\left\{\frac{K_h(x_j, x_i)}{\sum_{k\in B} K_h(x_j, x_k)}\right\}$$

and $n = \min(n_A, n_B)$. In addition, we have

$$E(\tilde{\theta}_{MIE} - \theta_N) = 0, \tag{7}$$

and

$$V(\tilde{\theta}_{MIE} - \theta_N) = V_A + V_B, \tag{8}$$

where

$$V_A = \text{plim}_{n,N\to\infty}\left\{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i\pi_j)\frac{m(x_i)}{\pi_i}\frac{m(x_j)}{\pi_j}\right\},$$

$$V_B = E\left[\frac{1}{N^2}\sum_{i\in B}\{g_B(x_i)\}^2\{y_i - m(x_i)\}^2\right].$$

According to Theorem 1, a consistent variance estimator of $\widehat{\theta}$ is

$$\widehat{V}_{np} = \frac{1}{N^2}\sum_{i\in A}\sum_{j\in A}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\frac{\widehat{m}(x_i)}{\pi_i}\frac{\widehat{m}(x_j)}{\pi_j} + \frac{1}{N^2}\sum_{i\in B}\{\widehat{g}_B(x_i)\}^2\widehat{e}_i^2, \tag{9}$$

where $\widehat{e}_i = y_i - \widehat{m}(x_i)$ and

$$\widehat{g}_B(x_i) = \sum_{j\in A} w_j\left\{\frac{K_h(x_j, x_i)}{\sum_{k\in B} K_h(x_j, x_k)}\right\}, \tag{10}$$

which is a kernel-based estimator of $\{\pi_B(x)\}^{-1}$, where $\pi_B(x) = \Pr(\delta = 1|x)$. Instead of using (10), one could consider using $\widehat{g}_B(x) = \{\widehat{\pi}_B(x)\}^{-1}$, in (9), where $\widehat{\pi}_B(x)$ is a kernel-based estimator of $E(\delta|x)$. This choice would require additional bandwidth selection and lead to more unstable results, so we do not pursue it further.

Remark 3.1. Instead of the linearization method, we can use a replication method for variance estimation of the nonparametric mass imputation estimator in (3). As an example, we use bootstrap. We first treat sample B as a simple random sample to obtain $\widehat{m}^{(k)}(x)$, for each $k = 1, \ldots, L$, where

$$\widehat{m}^{(k)}(x) = \frac{\sum_{j\in B} K_h(x, x_j^{*(k)}) y_j^{*(k)}}{\sum_{j\in B} K_h(x, x_j^{*(k)})}, \tag{11}$$

where $(x_j^{*(k)}, y_j^{*(k)})$ are the bootstrap resample of $\{(x_j, y_j) : j \in B\}$, under the "working" assumption that sample B is a simple random sample. The bootstrap sample size is the sample size of sample B. Once $\widehat{m}^{(k)}(\cdot)$ is obtained from (11), we can compute

$$\widehat{\theta}_{MIE}^{(k)} = \frac{1}{N}\sum_{i\in A} w_i^{(k)} \widehat{m}^{(k)}(x_i), \tag{12}$$

where $w_i^{(k)}$ is the bootstrap weight for w_i under the sampling design for sample A. The bootstrap variance estimator of $\widehat{\theta}_{MIE}$ is

$$\widehat{V}_{\text{boot}} = \frac{1}{L}\sum_{k=1}^{L}\left(\widehat{\theta}_{MIE}^{(k)} - \widehat{\theta}_{MIE}\right)^2.$$

By using techniques similar to those described in Kim et al. (2018), one can show that the above bootstrap variance estimator is consistent under certain regularity conditions.

Remark 3.2. The above kernel weighting framework covers k-nearest neighbor imputation as a special case by adopting a special kernel function. The k-nearest neighbor approach to mass imputation can be described in the following steps:

Step 1. For each unit i in sample A, find the k nearest neighbors from sample B, with the index set $\mathcal{J}_k(i) = \{i(1), \ldots, i(k)\}$ by using Euclidean distance based on the covariate x. Impute the y value for unit i by $\widehat{m}(x_i) = k^{-1}\sum_{j=1}^{k} y_{i(j)}$.

Step 2. The k-nearest neighbor imputation estimator of θ_N is

$$\widehat{\theta}_{\text{knn}} = \frac{1}{N}\sum_{i\in A} w_i \widehat{m}(x_i).$$

To see the connection between $\widehat{\theta}_{\text{knn}}$ and the kernel weighting estimator, we re-express

$$\widehat{m}(x) = \frac{\sum_{j\in B} K_{R_x}(x - x_j) y_j}{\sum_{j\in B} K_{R_x}(x - x_j)},$$

where

$$K_h(u) = \frac{1}{h^p} K\left(\frac{u}{h}\right), \; K(u) = 0.5I(||u|| \leq 1),$$

and the bandwidth $h = R_x$ is the random distance between x and its furthest among the k nearest neighbors. Therefore, $\widehat{\theta}_{\text{knn}}$ can be viewed as a kernel estimator incorporating a data-driven bandwidth.

4. GENERALIZED ADDITIVE MODELING

To overcome the curse of dimensionality of the kernel smoothing approach, we consider using GAM (Hastie and Tibshirani 1990) for mass imputation.

In GAM, we assume that y_i given $x_i = (x_{1,i}, \ldots, x_{p,i})$ follows an exponential family distribution with

$$q^{-1}\{m(x_i)\} = f_1(x_{1,i}) + f_2(x_{2,i}) + \cdots f_p(x_{p,i}), \tag{13}$$

where $q(\cdot)$ is an inverse link function (see two examples below), and each $f_k(\cdot)$ is a smooth function of $x_{k,i}$, for $k = 1, \ldots, p$. Because the function $f_k(x_k)$ is not restricted to a linear relationship of y and x_k, (13) specifies a flexible specification of the dependence of y on x.

Example 4.1. *For a continuous outcome, the Gaussian density function is*

$$f(y|x) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left[-\{y - m(x)\}/(2\sigma_y^2)\right],$$

where $q(\cdot)$ in (13) is an identity function.

Example 4.2. *For a binary outcome, the logistic density function is*

$$f(y|x) = \{m(x_i)\}^{y_i}\{1 - m(x_i)\}^{1-y_i},$$

where $q(\cdot)$ in (13) is a logit function.

There are several challenges in fitting model (13). First, $f_k(v)$ is an infinite-dimensional parameter, estimation of which often relies on some approximation. Second, we need to decide how smooth $f_k(v)$ should be to achieve the bias-variance trade-off in the estimation stage.

A common way to resolve the first challenge is to approximate $f_k(v)$ by splines. More specifically, let $B_m(v)$ be the basis spline functions for $m = 1, \ldots, M$ (Ruppert, Wand, and Carroll 2009). We approximate $f_k(v)$ by $f_k(v)$

$= \sum_{m=1}^{M} \gamma_m^k B_m(v)$ with spline coefficients γ_m^k. This leads to an approximation of model (13):

$$q^{-1}\{\widehat{m}(x_i)\} = \sum_{k=1}^{p} \sum_{m=1}^{M} \gamma_m^k B_m(x_{k,i}). \tag{14}$$

A large value of M allows for increased model complexity and an increased chance of overfitting; a small M may result in an inadequate model. One strategy is to choose a relatively large M and then penalize the model complexity in the estimation stage (Eilers and Marx 1996). Let the vector of spline coefficients be $\gamma_k^{\mathrm{T}} = (\gamma_1^k, \ldots, \gamma_M^k)$ and $\gamma^{\mathrm{T}} = (\gamma_1^{\mathrm{T}}, \ldots, \gamma_p^{\mathrm{T}})$. The corresponding likelihood function of γ is

$$L(\gamma) = \prod_{i \in B} f(y_i | x_i; \gamma), \tag{15}$$

where $f(y|x;\gamma)$ is the conditional density function of an exponential family distribution, see Examples 4.1 and 4.2, with (14). The estimate $\widehat{\gamma}$ is obtained by maximizing the penalized likelihood:

$$-2l(\gamma) + \sum_{k=1}^{p} \lambda_k \gamma_k^{\mathrm{T}} S_k \gamma_k, \tag{16}$$

where $l(\gamma) = \log L(\gamma)$ is the log-likelihood function of γ, S_k is a matrix with the (m, l)th component $\int B_m{}''(v) B_l{}'(v) \mathrm{d}v$, and $\gamma_k^{\mathrm{T}} S_k \gamma_k$ regularizes f_k to be smooth for which the degree of smoothness is controlled by λ_k. Given the smoothing parameter $\lambda^{\mathrm{T}} = (\lambda_1, \ldots, \lambda_p)$, the penalized likelihood function in (16) is optimized by a penalized version of the iteratively reweighted least squares algorithm (Nelder and Baker 1972) to obtain $\widehat{\gamma}$.

Regarding the choice of λ, we note that λ controls the trade-off between model complexity and overfitting, which can be estimated separately from other model coefficients using generalized cross-validation or estimated simultaneously using restricted maximum likelihood estimation (Wood 2006).

Once the model is fitted, we can create an imputed value for each element i in Sample A as

$$\widehat{m}_{\mathrm{gam}}(x_i) = q\left\{ \sum_{k=1}^{p} \sum_{m=1}^{M} \widehat{\gamma}_m^k B_m(x_{i,k}) \right\}.$$

The mass imputation estimator based on the GAM is

$$\widehat{\theta}_{\mathrm{gam}} = \frac{1}{N} \sum_{i \in A} w_i \widehat{m}_{\mathrm{gam}}(x_i).$$

4.1. Hybrid Bootstrap for Variance Estimation

Variance estimation of $\widehat{\theta}_{\text{gam}}$ is challenging because the penalty term in (16) regularizes the variance of $\widehat{\gamma}$ at the expense of introducing a bias. Therefore, we expect that the linearization method for variance estimation would not work well for $\widehat{\theta}_{\text{gam}}$. We propose a hybrid method that is obtained by combining bootstrap and Bayesian inference. From a Bayesian framework, we have the posterior distribution of γ as

$$\gamma|\text{data} \sim N\{\widehat{\gamma}, (B^{\text{T}}WB + \sum_{k=1}^{p} \lambda_k S_k)^{-1}\sigma^2\}, \tag{17}$$

asymptotically, where B is the matrix with the $i$the row $\{B_1(x_{1,i}), \ldots, B_M(x_{1,i}), \ldots, B_1(x_{k,i}), \ldots, B_M(x_{k,i})\}$, W is a diagonal matrix with $W_{ii} = [\dot{q}\{m(x_i)\}^2 V\{m(x_i)\}]^{-1}$ and $\dot{q}(x) = \text{d}q(x)/\text{d}x$; see Wood (2006). We propose a replication method for variance estimation of $\widehat{\theta}_{\text{gam}}$. In particular, we use the bootstrap method as follows:

Step 1. For each $b = 1, \ldots, L$, we sample $\gamma^{*(b)}$ from (17) and obtain obtain $\widehat{m}_{\text{gam}}^{(b)}(x)$, where

$$\widehat{m}_{\text{gam}}^{(b)}(x) = g\{\sum_{k=1}^{p}\sum_{m=1}^{M} \widehat{\gamma}_m^{*(b)k} B_m(x_{i,k})\}. \tag{18}$$

The sampling $\gamma^{*(b)}$ from (17) can be implemented easily from the "mgcv" package in R, see Wood (2019).

Step 2. We can compute

$$\widehat{\theta}_{\text{gam}}^{(b)} = \frac{1}{N}\sum_{i\in A} w_i^{(b)} \widehat{m}_{\text{gam}}^{(b)}(x_i), \tag{19}$$

where $w_i^{(k)}$ is the bootstrap weight for w_i under the sampling design for sample A.

The bootstrap variance estimator of $\widehat{\theta}_{\text{gam}}$ is then obtained by

$$\widehat{V}_{\text{gam}} = \frac{1}{L}\sum_{b=1}^{L}\left(\widehat{\theta}_{\text{gam}}^{(b)} - \widehat{\theta}_{\text{gam}}\right)^2.$$

Instead of the hybrid approach of combining bootstrap and Bayesian inference, one can also develop a fully Bayesian approach, which is covered in the

following section.

4.2. Approximate Bayesian Method for Variance Estimation

We introduce an approximate Bayesian approach to data integration, under the setup of Section 2. To fix ideas, we first consider a mass imputation using parametric model $f(y|x;\alpha)$ first and then discuss nonparametric imputation. Under a parametric model $f(y|x;\alpha)$ with a prior distribution on α as $\pi(\alpha)$, a posterior distribution of α can be obtained from sample B as follows:

$$p(\alpha|\text{data}_B) = \frac{L_B(\alpha)\pi(\alpha)}{\int L_B(\alpha)\pi(\alpha)d\alpha}, \tag{20}$$

where $L_B(\alpha) = \coprod_{i\in B} f(y_i|x_i;\alpha)$ is the likelihood function of α from sample B and $\text{data}_B = \{(x_i, y_i), i \in B\}$. Using the posterior distribution in (20), we can create mass imputation from the posterior predictive distribution.

$$p(y_i|x_i, \text{data}_B) = \int f(y_i|x_i;\alpha)p(\alpha|\text{data}_B)d\alpha.$$

Now, if y_i were observed from sample A, using the idea of Wang, Kim, and Yang (2018), the posterior distribution of θ_N with the prior $\pi(\theta)$ would be approximately computed by

$$p(\theta|\widehat{\theta}_A) = \frac{g(\widehat{\theta}_A|\theta)\pi(\theta)}{\int g(\widehat{\theta}_A|\theta)\pi(\theta)}, \tag{21}$$

where $g(\widehat{\theta}_A|\theta)$ is the density of the sampling distribution of $\widehat{\theta}_A = N^{-1}\sum_{i\in A} w_i y_i$ and is often approximated by

$$\frac{\widehat{\theta}_A - \theta}{\{\widehat{V}(\widehat{\theta}_A)\}^{1/2}} \xrightarrow{\mathcal{L}} N(0,1),$$

where $\widehat{V}(\widehat{\theta}_A)$ is the design-consistent variance estimator of $\widehat{\theta}_A$.

The proposed Bayesian approach for parametric mass imputation can be summarized as follows:

Step 1. Generate M posterior values of α, denoted by $\alpha^{*(1)}, \ldots, \alpha^{*(M)}$, from (20).

Step 2. For each $j = 1, \ldots, M$, generate $y_i^{*(j)}$ from $f(y_i|x_i;\alpha^{*(j)})$ for all $i \in A$.

Step 3. Using $y_i^{*(j)}$ generated from Step 2, generate the posterior value of θ from (21) with $\widehat{\theta}_A$ replaced by $\widehat{\theta}_I^{*(j)} = N^{-1}\sum_{i\in A} w_i y_i^{*(j)}$ and $\widehat{V}(\widehat{\theta}_A)$ computed

by the j-th imputed data. We can use the M posterior values of θ to make Bayesian inference about θ.

For the nonparametric mass imputation using GAM, we must change the posterior step using (17). Once the parameters for GAM are generated from the posterior step, we can use $f(y|x;\gamma^*)$ to obtain the imputed values, where $f(y|x;\gamma)$ is defined in (15). (Step 3) remains the same.

5. SIMULATION STUDIES

5.1 Simulation Study One

In this section, we assess the finite-sample performance of the proposed estimators via simulation from two artificial statistical models. We use the following models to generate two finite populations of size $N = 10{,}000$.

(1) Model I: The y_i's are independently generated from $N(0.3 + 2x_{1i} + 2x_{2i}, 1)$, where $x_{1i} \overset{i.i.d}{\sim} N(2,1)$ and $x_{2i} \overset{i.i.d}{\sim} N(2,1)$.
(2) Model II: The y_i's are independently generated from $N(0.3 + 0.5x_{1i}^2 + 0.5x_{2i}^2, 1)$, where $x_{1i} \overset{i.i.d}{\sim} N(2,1)$ and $x_{2i} \overset{i.i.d}{\sim} N(2,1)$.

From each of the two finite populations, we generate two independent samples. We use simple random sampling of size $n_A = 500$ to obtain sample A. In selecting sample B of size $n_B = 500$ and 1,000, we create two strata, where Stratum 1 consists of elements with $x_{1i} \leq 2$ and Stratum 2 consists of elements with $x_{1i} > 2$. The population size for each of the two strata is about 5,000. Within each stratum $t \in \{1,2\}$, we select n_t units by simple random sampling, independently between the two strata, where $n_1 = 0.7n_B$ and $n_2 = 0.3n_B$. We assume that the stratum information is unavailable at the time of data analysis. Thus, the sampling mechanism for sample B is unknown for data analysis, but it satisfies the noninformativeness assumption. Using the two samples, we compute six estimators of finite population mean $\theta_{1N} = N^{-1}\sum_{i=1}^{N} y_i$ and finite population domain mean $\theta_{2N} = \{\sum_{i=1}^{N} I(x_{1i} > 2)\}^{-1} \sum_{i=1}^{N} y_i I(x_{1i} > 2)$. The six estimators are listed as follows:

(1) The sample mean from sample A (Mean A): $\widehat{\theta}_A = n_A^{-1}\sum_{i\in A} y_i$.
(2) The naive estimator (sample mean) from sample B (Mean B): $\widehat{\theta}_B = n_B^{-1}\sum_{i\in B} y_i$.
(3) The parametric mass imputation estimator (PMIE): $\widehat{\theta}_{\text{PMIE}} = n_A^{-1}\sum_{i\in A}\widehat{y}_i$ with $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i}$, where $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)$ are the estimated regression coefficients obtained from sample B.
(4) The pseudo-weighted estimator (PWE) proposed by Elliott and Valliant (2017).
(5) The proposed nonparametric mass imputation estimator (NPMIEK) using a Kernel regression in (4). Both x_1 and x_2 are used as the predictors in

kernel smoothing function. Multivariate Gaussian kernel density function is used.

(6) The proposed nonparametric mass imputation estimator (NPMIEG) by using a GAMs in (13).

In computing NPMIEK, the optimal bandwidth is selected by 10-fold cross-validation from sample B, which minimizes the mean squared prediction error. The sample mean of sample A serves as a gold standard estimator. Results are based on 1,000 repeated simulation runs. Table 2 presents the Monte Carlo relative bias (RB), relative standard error (RSE), and relative root mean squared error of the six point estimators for estimating population mean and domain mean.

The Mean A estimator is a gold standard, but it is not applicable in the setup of data integration. It is used as a benchmark for comparison. For population mean estimation, the Mean B estimator is seriously biased due to the sampling mechanism for sample B. PMIE shows good performance in Model I, as y is linearly related to x_1 and x_2. However, in Model II where the linear relationship among x_1, x_2, and y fails, the PMIE is slightly biased, and the bias does not decrease with an increased sample size for sample B. The PWE has small RBs under both models, but it has larger RSEs and relative root mean square errors (RRMSEs) than do other estimators, besides the Mean B estimator. The NPMIEK and NPMIEG are also biased modestly in both models, but the biases decrease as the sample size for sample B increases. NPMIEK shows slightly larger biases than does NPMIEG in both models. In terms of RRMSE, the PMIE and NPMIEG are the smallest under model I because the linear model is correctly specified. Under model II, the NPMIEG has smaller RRMSEs than does the PMIE, as its biases are smaller. The NPMIEK has smaller biases than does the PMIE and RSEs and RRMSEs that are comparable to those of the PMIE. Note that the mass imputation estimator can have a smaller MSE than the gold standard. This possibility can be explained by the fact that the mass imputation estimator can effectively incorporate all of the available information from both samples. In this simulation, NPMIEG performs slightly better than NPMIEK. For domain mean estimation, the Mean B estimator shows small RBs, since within the stratum the sampling process for sample B is simple random sampling and the pseudo-inclusion probabilities are all equal in stratum 2. Our proposed estimators NPMIEK and NPMIEG show better performance in terms of balancing RBs and RSEs than do PMIE and PWE. The PWE method shows big RBs, which is consistent with the results in Chen et al. (2019).

We also compute the proposed bootstrap variance estimator and the corresponding confidence intervals with 95 percent nominal coverage rates (CRs). Table 3 presents the performance of the variance estimators and confidence intervals. The proposed variance estimator shows negligible RBs for all scenarios, and the coverage probabilities for our proposed methods are close to the nominal rate in all scenarios. Under Model I, the CRs for the PMIE method are

Table 2. Monte Carlo RB, Monte Carlo RSE, and RRMSE of the Six Point Estimators of Population Mean and Domain Mean, Based on 1,000 Monte Carlo Samples

		Model I			Model II		
n_B	Estimator	RB (%)	RSE (%)	RRMSE (%)	RB (%)	RSE (%)	RRMSE (%)
Population mean estimation							
	Mean *A*	0.03	1.63	1.63	−0.01	2.65	2.65
	Mean *B*	−7.65	1.35	7.77	−12.07	2.27	12.28
500	PMIE	0.00	1.64	1.64	−1.10	2.64	2.86
	PWE	−0.01	1.74	1.74	−0.01	3.17	3.17
	NPMIEK	−0.15	1.66	1.67	−1.08	2.65	2.86
	NPMIEG	0.00	1.65	1.65	−0.08	2.63	2.63
	Mean *A*	0.01	1.59	1.59	0.01	2.72	2.72
	Mean *B*	−7.67	0.98	7.73	−12.04	1.55	12.14
1000	PMIE	0.00	1.56	1.56	−1.07	2.50	2.72
	PWE	0.05	1.62	1.62	0.14	2.81	2.82
	NPMIEK	−0.10	1.57	1.57	−0.68	2.66	2.75
	NPMIEG	0.00	1.56	1.56	−0.02	2.66	2.66
Domain mean estimation							
	Mean *A*	−0.03	1.67	1.67	0.00	2.76	2.76
	Mean *B*	−0.03	2.03	2.03	0.00	3.63	3.63
500	PMIE	−0.06	1.72	1.72	−2.86	2.78	3.99
	PWE	2.65	2.09	3.37	6.19	4.03	7.38
	NPMIEK	−0.51	1.79	1.86	−1.47	2.86	3.22
	NPMIEG	−0.06	1.72	1.72	−0.08	2.79	2.79
	Mean *A*	0.04	1.64	1.64	0.00	2.84	2.84
	Mean *B*	0.02	1.44	1.44	−0.02	2.47	2.47
1000	PMIE	0.02	1.57	1.57	−2.84	2.49	3.78
	PWE	2.77	1.65	3.22	6.25	2.87	6.88
	NPMIEK	−0.28	1.61	1.63	−0.93	2.77	2.92
	NPMIEG	0.02	1.57	1.57	0.00	2.75	2.75

comparable with the results for our proposed methods. Under Model II, the CRs for the PMIE method are lower than our proposed methods. The CR for the PMIE method for domain estimation is extremely low since the model assumptions are violated. NPMIEG performs slightly better than NPMIEK because of its smaller biases of the point estimators.

Table 3. Monte Carlo RB of Standard Error Estimators, CRs of 95 Percent Confidence Intervals, and ALs of the Confidence Intervals of Population Mean and Domain Mean, Based on 1,000 Monte Carlo Samples with a Sample of Size 500

		Model I			Model II		
Parameter	Estimator	RB (%)	CR (%)	AL (%)	RB (%)	CR (%)	AL (%)
Mean	PMIE	−3.22	93.7	51.6	−1.62	92.9	53.9
	NPMIEK	−3.43	93.7	52.1	1.03	93.5	55.5
	NPMIEG	−3.07	94.2	52.0	−0.09	94.5	54.5
Domain mean	PMIE	−6.53	93.6	62.3	−1.90	81.0	73.6
	NPMIEK	−1.97	93.0	67.8	2.65	91.0	79.0
	NPMIEG	−3.56	94.5	64.2	−0.91	94.0	74.6

5.2 Simulation Study Two

In this section, we conduct another Monte Carlo simulation study based on 2017–2018 US National Health Nutrition and Examination Survey (NHANES) data. The 2017–2018 NHANES is a stratified multistage household survey that oversampled certain minority groups, including Hispanic or Black and Asian populations. The target population is the noninstitutionalized civilian population, including all people living in households and excluding institutional group quarters and those persons on active duty with the military. The primary objective of NHANES is to produce a broad range of descriptive health and nutrition statistics for sex, race and Hispanic origin, and age group of the US population. For more information about the design and objectives of NHANES, please see https://wwwn.cdc.gov/nchs/nhanes/analyticguidelines.aspx/sample-design.

In this simulation study, we first create a subset of the original 2017–2018 NHANES data by removing cases with missing values on the following variables: age, body mass index (BMI), and diastolic blood pressure. This resulted in a data file of 6,230 cases. For simplicity, we treat the subset sample as the finite population for Monte Carlo simulation. We consider Diastolic blood pressure as the outcome variable and Age and BMI as predictors. The scatter plots of Diastolic blood pressure versus Age and Diastolic blood pressure versus BMI are presented in figure 1. There seems to be a nonlinear relationship between the outcome variable and predictors. We consider a Monte Carlo sample size of $B = 1{,}000$. For each Monte Carlo sample, we generate two independent samples. We use simple random sampling of size $n_A = 500$ to obtain sample A. In selecting sample B of size $n_B = 500$, we create two strata, where Stratum 1 consists of elements with Age less than or equal to 40 and Stratum 2 consists of elements with Age greater than 40. The population sizes for the two strata

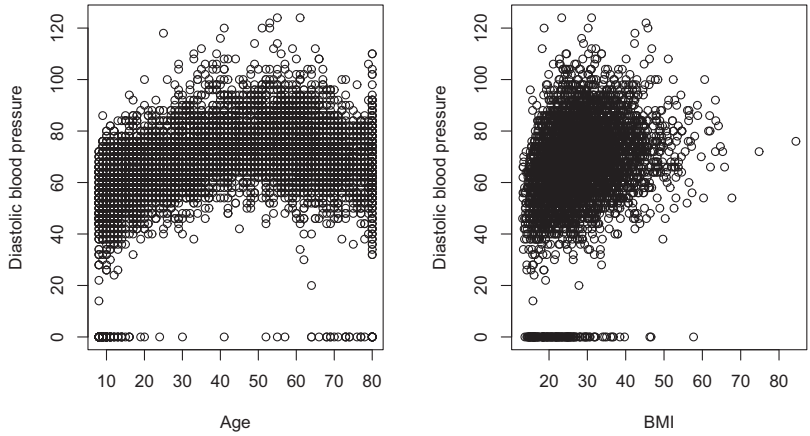

Figure 1. Scatter plot of study variable versus predictors.

are 3,015 and 3,082. Within each stratum $t \in \{1, 2\}$, we select n_t units by simple random sampling, independent between the two strata, where $n_1 = 0.7n_B$ and $n_2 = 0.3n_B$. We assume that the stratum information is unavailable at the time of data analysis. Using the two samples, we compute six estimators considered in simulation study one for estimating the finite population mean of Diastolic blood pressure and the corresponding domain mean for cases with Age greater than 40.

The results of point estimation are presented in table 4. The Mean *A* estimator performs the best in both scenarios since it is the gold standard and assumes that one observes study variable in probability sample *A*. For estimating the population mean, the Mean *B* estimator shows a large RB because it is only based on sample *B*. NPMIEK and NPMIEG show smaller RB than do PMIE and PWE, since the underlying relationship between the study variable and predictors is not linear. NPMIEK, NPMIEG, PMIE, and PWE have comparable RSE and RRMSE. For estimating domain mean, the Mean *B* estimator shows small bias, since the sampling design within the domain is simple random sampling and the pseudo-inclusion probabilities are all equal in stratum 2. PMIE and PWE show greater RB, RSE, and RRMSE than do NPMIEK and NPMIEG.

In addition, we also compare the performance of our proposed bootstrap variance estimators in terms of RB, CR, and average lengths (ALs) with PMIE. The results are presented in table 5. All RBs are small (less than 2 percent in terms of absolute value). For both scenarios, NPMIEK and NPMIEG show better CRs than does PMIE. PMIE shows very low CR, especially for estimating the domain mean.

Table 4. Monte Carlo RB, Monte Carlo RSE, and RRMSE of the Six Point Estimators of Population Mean and Domain Mean, Based on 1,000 Monte Carlo Samples and NHANES Data

Parameter	Estimator	RB (%)	RSE (%)	RRMSE (%)
	Mean *A*	−0.04	1.02	1.03
	Mean *B*	−3.04	1.02	3.21
Mean	PMIE	0.29	1.17	1.20
	PWE	−0.22	1.12	1.14
	NPMIEK	0.07	1.12	1.12
	NPMIEG	−0.01	1.16	1.16
	Mean *A*	0.01	1.19	1.19
	Mean *B*	−0.03	1.59	1.59
Domain mean	PMIE	1.03	1.72	2.01
	PWE	−1.19	1.72	2.09
	NPMIEK	−0.06	1.53	1.53
	NPMIEG	0.07	1.57	1.58

Table 5. Monte Carlo RB of Standard Error Estimators, CRs of 95 Percent Confidence Intervals, and ALs of the Confidence Intervals of Population Mean and Domain Mean, Based on 1,000 Monte Carlo Samples with Sample Size 500 and NHANES Data

Parameter	Estimator	RB (%)	CR (%)	AL
	PMIE	0.93	94.2	3.13
Mean	NPMIEK	−0.15	95.0	2.96
	NPMIEG	1.29	95.5	3.11
	PMIE	0.74	89.8	4.92
Domain mean	NPMIEK	−1.69	94.1	4.26
	NPMIEG	1.71	95.1	4.53

6. REAL DATA APPLICATION

We consider a real data application using a probability sample from the KNHANES and a nonprobability sample from NHISS. The KNHANES is a national survey that studies the health and nutritional status of Koreans and has been conducted annually since 1998. Both surveys are conducted by the Korea Centers for Disease Control and Prevention. The KNHANES is a nationally representative cross-sectional survey that includes approximately 10,000 individuals each year as a survey sample and collects information on social-economic status, health-related behaviors, quality of life, healthcare utilization, anthropometric measures, biochemical and clinical profiles for

noncommunicable diseases, and dietary intakes with three component surveys: health interview, health examination, and nutrition survey. More details of the KNHANES can be found in Kweon, Kim, Jang, Kim, Kim, et al. (2014). The nonprobability sample from NHISS provides health-related information collected from National Health Screening Program (NHSP) in South Korea. The NHSP was launched with the goal of improving the overall health of the South Korean citizens and preventing costly chronic diseases. All beneficiaries are eligible for screening once every year or two, depending on their demographic or occupational status. The specific screening items are stipulated by the implementation standards, which include various blood tests and cancer screening. The total number of eligible beneficiaries is about 16 million, and approximately 75 percent participated in the screening. The data that we used in the present study are from the subset corresponding to the blood test results that are associated with metabolic syndrome from the 2014 program. The variables in this data set are demographics, such as sex and age, and clinical measurements, such as total glycerides (mg/dL), total cholesterol (mg/dL), high-density lipoprotein cholesterol (HDL, mg/dL), and medical diagnosis of anemia. The data set is made publicly available after anonymization and random selection of 1 million observations (National Health Insurance Data Sharing Service, https://nhiss.nhis.or.kr/bd/ab/bdabf006cv.do). More thorough data can be purchased with a paid subscription and expert panel review.

In our real-world application, we treat the KNHANES subsample data for blood test as the probability sample A with sample size 4,929 after removing the missing values for key items. To reduce the computational burden, we first draw a simple random sample with size 5,000 from the original NHISS data and treat the subsample as the nonprobability sample B for our analysis. We consider the following variables as predictors: Sex, Age, Hemoglobin (HGB), Triglyceride (TG), and High-density Lipoprotein Cholesterol (HDL, mg/dL). Age has 27 categories as following: 1 for "20 to 24," 2 for "25 to 26," 3 for ":27 to 28,". . .,27 for "75 or higher." Total Cholesterol (TCHOL) is considered the outcome variable of interest. TCHOL is a variable in both KNHANES and NHISS, so we can evaluate the performance of different approaches by comparing the estimates with weighted estimates calculated from KNHANES. The estimated means of predictors from sample KNHANES and sample NHISS data are presented in table 6. There is some discrepancy between the two samples. For example, the estimated prevalence of "Male" from KNHANES is about 42.7 percent, while the estimated prevalence of "Male" is 51.7 percent for NHISS. The parameters of interest are the population mean of TCHOL and the domain mean of TCHOL for the "Male" group.

We consider the six estimators presented in Section 5 and use weighted mean with probability sample A (Mean A) as the benchmark to calculate the Biases. For NPMIEK, we use kernel smoothing-based mass imputation for "Male" and "Female" groups separately. Within each group, Age, HGB, TG, and HDL are used as the predictors in kernel density function. Multivariate

Table 6. Estimated Population Means of Predictors from KNHANES and NHISS

Covariates	KNHANES	NHISS
Sex (male%)	42.7	51.7
Age	14.4	14.0
HGB	14.1	14.1
TG	136.4	130.5
HDL	51.1	55.2

Table 7. RBs (Percent) of Estimated Population Mean and Domain Mean for TCHOL (the Bootstrap Standard Errors are within Parentheses)

Parameter	Mean *B*	PMIE	PWE	NPMIEK	NPMIEG
Mean	4.04	3.33 (0.63)	3.10	2.14 (0.82)	2.20 (0.64)
Domain mean	3.95	4.26 (0.86)	3.17	2.77 (1.06)	2.62 (0.85)

Gaussian kernel density function is used. We also calculate standard errors for mass imputation estimators PMIE, NPMIEK, and NPMIEG. The results are presented in table 7. As shown in table 7, our proposed nonparametric mass imputation estimators NPMIEK and NPMIEG have smaller RBs than do other estimators. The boostrap standard errors for NPMIEG are similar to those of PMIE, and NPMIEK has slightly larger standard errors.

7. DISCUSSION

As demonstrated in this paper, mass imputation is a promising tool for data integration, as long as the same imputation model holds for both the probability sample and the nonprobability sample. The mass imputation estimator is able to reduce the selection bias of the naive estimator, which is based solely on the nonprobability sample. In the simulation studies, we use domains that correspond exactly with one of the implicit sampling strata for the sample *B*. Note that domains that cut across implicit sampling strata in sample *B* are also typically of analytic interest in practice and may produce different results from those in our simulations. In practice, more complex models with many covariates can be implemented for estimating the pseudo-inclusion probabilities in Elliott and Valliant (2017). To determine the predictors for our proposed nonparametric modeling process, one can first conduct univariate analysis and only select predictors that are significantly correlated with the study variables of interest. Then, colinearity analysis can be performed to further reduce the dimension of predictors. Furthermore, dimension reduction techniques, such as a single index model, can also be used. Moreover, by implementing non/semiparametric models, we mitigate the potential bias due to parametric model misspecification. However, there is no guarantee that the noninformative sampling or

transportability assumption holds in practice. In this case, one possible solution is to obtain a validation subsample from the original probability sample and collect information for the study variable of interest to assess the validity of the mass imputation estimator. Use of a validation subsample is essentially a two-phase sampling problem and involves extra costs. Another option is to develop an explicit model for the selection mechanism for the nonprobability sample and develop mass imputation even when the noninformative sampling assumption does not hold. Model specification and parameter estimation under nonignorable nonresponse models can be used in this scenario. These extensions will be presented elsewhere. In practice, instead of using our proposed nonparametric mass imputation approaches, one can also use other machine learning-based mass imputation approaches, such as regression trees or random forests. The machine learning-based approaches may work better with more complex model structures with many interaction terms, for instance. The comparison of our proposed methods with such methods will be an interesting future research topic.

APPENDIX

A.1: Regularity Conditions

We specify regularity conditions for Theorem 1 here.

(C1) $f(x)$ and $\pi_B(x)$ have bounded partial derivatives with respect to x up to an order t with $t \geq 2$, $2t > d_x$ almost surely, where $f(x)$ is the density of x, $\pi_B(x) = \Pr(r = 1|x)$ and d_x is the dimension of x.

(C2) The kernel function $K(s)$ is a probability density function such that

(a) It is bounded and has compact support.

(b) $\int K(s_1, \ldots, s_{d_x}) ds_1 \ldots ds_{dx} = 1$.

(c) $\int s_i^l K(s_1, \ldots, s_{d_x}) ds_1 \ldots ds_{d_x} = 0$ for any $i = 1, \ldots, d_x$ and $1 \leq l < q$.

(d) $\int s_i^q K(s_1, \ldots, s_{d_x}) ds_1 \ldots ds_{d_x} \neq 0$.

(C3) $nh^{2d_x} \to \infty$, $n^{1/2} h^q \to 0$, as $n \to \infty$ and $N \to \infty$.

(C4) $1 > \pi_B(x) > d > 0$ almost surely.

(C5) $E\{m^2(x)\} < \infty$ and $m(x)$ has continuous partial derivative $\partial m(x)/\partial x$ and $E\{\partial m(x)/\partial x\} < \infty$.

(C6) $\pi_B(x)$ has continuous partial derivative $\partial \pi_B(x)/\partial x$ with $E\{\partial \pi_B(x)/\partial x\} < \infty$.

Conditions (C1)–(C3) are common conditions used for nonparametric problems, see also (Wang and Chen 2009). $n^{1/2} h^q \to 0$ is used in condition (C3) to control the bias due to kernel smoothing, and $nh^{2d_x} \to \infty$ is used to produce a consistent estimator of the conditional distribution and control the convergence rate of response probability estimation. Condition (C4) is used to avoid extreme propensity scores. Conditions (C5) and (C6) are standard conditions to control the moments and continuity.

A.2: Sketched Proof of THEOREM 1

For simplicity, we assume $q = 2$ and $d_x = 1$ in the following proof. A similar proof can be obtained for other cases. Define $\eta(x) = \pi_B(x)f(x)$ and $\widehat{\eta}(x) = N^{-1}\sum_{j\in B}K_h(x_i, x_j)$ as its kernel estimator with $K_h(x_i, x_j) = h^{-d_x}K\{h^{-1}(x_i - x_j)\}$, then we have

$$\begin{aligned}
\widehat{\theta}_{MIE} - \theta_N &= \frac{1}{N}\sum_{i\in A} w_i\widehat{m}(x_i) - \theta_N \\
&= \frac{1}{N}\sum_{i\in A} w_i m(x_i) - \theta_N + \frac{1}{N}\sum_{i\in A} w_i\{\widehat{m}(x_i) - m(x_i)\} \\
&= \frac{1}{N}\sum_{i\in A} w_i m(x_i) - \theta_N + \frac{1}{N}\sum_{i\in A} w_i \frac{N^{-1}\sum_{j\in B}K_h(x_i, x_j)\{y_j - m(x_j)\}}{\eta(x_i)} \\
&\quad + \frac{1}{N}\sum_{i\in A} w_i\{\widehat{m}(x_i) - m(x_i)\}\frac{\eta(x_i) - \widehat{\eta}(x_i)}{\eta(x_i)} \\
&\quad + \frac{1}{N}\sum_{i\in A} w_i \frac{N^{-1}\sum_{j\in B}K_h(x_i, x_j)\{m(x_j) - m(x_i)\}}{\eta(x_i)} \\
&= T_1 + T_2 + T_3 + T_4.
\end{aligned} \tag{A.2.1}$$

Define $\zeta_{ij} = w_i I_i \delta_j K_h(x_i, x_j)\{y_j - m(x_j)\}\eta^{-1}(x_i)$, then we have

$$\begin{aligned}
T_2 &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_i I_i \delta_j K_h(x_i, x_j)\{y_j - m(x_j)\}\eta^{-1}(x_i) \\
&= \frac{1}{N(N-1)}\sum_{i\neq j} H(z_i, z_j) + o_p(n^{-1/2}),
\end{aligned} \tag{A.2.2}$$

where $z_i = (x_i, y_i, \delta_i, I_i)$ and

$$\begin{aligned}
H(z_i, z_j) &= \frac{1}{2}[w_i I_i \delta_j K_h(x_i, x_j)\frac{\{y_j - m(x_j)\}}{\eta(x_i)} \\
&\quad + w_j I_j \delta_i K_h(x_j, x_i)\frac{\{y_i - m(x_i)\}}{\eta(x_j)}] := \frac{1}{2}(\zeta_{ij} + \zeta_{ji}).
\end{aligned} \tag{A.2.3}$$

According to (B.2) and (B.3), we know that $\sum_{i\neq j}H(z_i, z_j)\{N(N-1)\}^{-1}$ is the U-statistic. Let $s = (x_j - x_i)h^{-1}$, by $nh^2 \to \infty$, $nh^4 \to 0$ and according to Taylor expansion, we have

$$E(\zeta_{ij}|z_i) = \frac{w_i I_i}{\eta(x_i)}E\Big\{\delta_j K\Big(\frac{x_j - x_i}{h}\Big)\{y_j - m(x_j)\}|z_i\Big\} = 0, \tag{A.2.4}$$

provided $E(y_j|x_j, \delta_j) = m(x_j)$, and

$$\begin{aligned} E(\zeta_{ji}|z_i) &= \delta_i\{y_i - m(x_i)\}E\{w_j I_j K_h(x_j, x_i)\eta^{-1}(x_j)|z_i\} \\ &= \delta_i\{y_i - m(x_i)\}E\Big\{h^{-1}K\Big(\frac{x_j - x_i}{h}\Big)\eta^{-1}(x_j)|z_i\Big\} \\ &= \delta_i\{y_i - m(x_i)\}h^{-1}\int K\Big(\frac{x_j - x_i}{h}\Big)\eta^{-1}(x_j)f(x_j)dx_j \\ &= \delta_i\{y_i - m(x_i)\}\int K(s)\eta^{-1}(x_i + hs)f(x_i + hs)ds \\ &= \delta_i\pi_B^{-1}(x_i)\{y_i - m(x_i)\} + O(h^2). \end{aligned} \tag{A.2.5}$$

According to (A.2.2)–(A.2.5) and by the theory of *U*-statistics, see (Serfling 1980), Chapter 5, we have

$$T_2 = \frac{1}{N}\sum_{i=1}^{N}\delta_i\pi_B^{-1}(x_i)\{y_i - m(x_i)\} + o_p(n^{-1/2}). \tag{A.2.6}$$

In addition, by using an argument similar to that in (Wang and Chen 2009), it can be shown that $T_3 = o_p(n^{-1/2})$ and $T_4 = o_p(n^{-1/2})$. Together with (A.2.1) and (A.2.6), we have

$$\widehat{\theta}_{MIE} - \theta_N = \frac{1}{N}\sum_{i\in A} w_i m(x_i) - \theta_N + \frac{1}{N}\sum_{i\in B}\pi_B^{-1}(x_i)\{y_i - m(x_i)\} + o_p(n^{-1/2}). \tag{A.2.7}$$

In addition, we have

$$\begin{aligned} \tilde{\theta}_{MIE} - \theta_N &= \frac{1}{N}\sum_{i\in A} w_i m(x_i) - \theta_N + \frac{1}{N}\sum_{j\in B} g_B(x_j)\{y_j - m(x_j)\} \\ &= \frac{1}{N}\sum_{i\in A} w_i m(x_i) - \theta_N + \frac{1}{N^2}\sum_{j\in B}\sum_{i=1}^{N}\Big\{\frac{K_h(x_i, x_j)}{N^{-1}\sum_{k\in B}K_h(x_i, x_k)}\Big\}\{y_j - m(x_j)\} \\ &= T_1 + \frac{1}{N}\sum_{i=1}^{N}\frac{N^{-1}\sum_{j\in B}K_h(x_i, x_j)\{y_j - m(x_j)\}}{\widehat{\eta}(x_i)} \\ &= T_1 + \frac{1}{N}\sum_{i=1}^{N}\frac{N^{-1}\sum_{j\in B}K_h(x_i, x_j)\{y_j - m(x_j)\}}{\eta(x_i)} \\ &\quad + \frac{1}{N}\sum_{i=1}^{N}\frac{N^{-1}\sum_{j\in B}K_h(x_i, x_j)\{y_j - m(x_j)\}\{\eta(x_i) - \widehat{\eta}(x_i)\}}{\widehat{\eta}(x_i)\eta(x_i)} \\ &= T_1 + T_2^* + T_3^*. \end{aligned} \tag{A.2.8}$$

By using techniques similar to those used in the proof of T_2 in (A.2.6), it can be shown that

$$T_2^* = \frac{1}{N}\sum_{i=1}^{N} \delta_i \pi_B^{-1}(x_i)\{y_i - m(x_i)\} + o_p(n^{-1/2}). \tag{A.2.9}$$

In addition, by using an argument similar to that in (Wang and Chen 2009), it can be shown that $T_3^* = o_p(n^{-1/2})$. Therefore, according to (A.2.8), (A.2.9), and (A.2.7), we have

$$\begin{aligned}\tilde{\theta}_{MIE} - \theta_N &= \frac{1}{N}\sum_{i\in A} w_i m(x_i) - \theta_N + \frac{1}{N}\sum_{i\in B} \pi_B^{-1}(x_i)\{y_i - m(x_i)\} + o_p(n^{-1/2}) \\ &= \widehat{\theta}_{MIE} - \theta_N + o_p(n^{-1/2}).\end{aligned} \tag{A.2.10}$$

Therefore, we have

$$\widehat{\theta}_{MIE} - \tilde{\theta}_{MIE} = o_p(n^{-1/2}).$$

REFERENCES

Breidt, F. J., A. McVey, and W. A. Fuller (1996). "Two-Phase Estimation by Imputation," *Journal of the Indian Society of Agricultural Statistics*, 49, 79–90.

Chen, Y., P. Li, and C. Wu (2019). "Doubly Robust Inference with Non-Probability Survey Samples," *Journal of the American Statistical Association*, doi: 10.1080/01621459.2019.1677241: https://www.tandfonline.com/doi/abs/10.1080/01621459.2019.1677241

Cheng, P. E. (1994). "Nonparametric Estimation of Mean Functionals with Data Missing at Random," *Journal of the American Statistical Association*, 89, 81–87.

Eilers, P. H., and B. D. Marx (1996). "Flexible Smoothing with B-Splines and Penalties," *Statistical Science*, 11, 89–102.

Elliott, M. R., and R. Valliant (2017). "Inference for Nonprobability Samples," *Statistical Science*, 32, 249–264.

Epanechnikov, V. A. (1969). "Non-Parametric Estimation of a Multivariate Probability Density," *Theory of Probability & Its Applications*, 14, 153–158.

Fuller, W. A. (2011). *Sampling Statistics*, vol. 560, John Wiley & Sons.

Hastie, T., and R. Tibshirani (1990). *Generalized Additive Models*, NY: Chapman and Hall, Inc.

Horvitz, D. G., and D. J. Thompson (1952). "A Generalization of Sampling without Replacement from a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

Keiding, N., and T. A. Louis (2016). "Perils and Potentials of Self-Selected Entry to Epidemiological Studies and Surveys," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179, 319–376.

Kim, J., S. Park, Y. Chen, and C. Wu (2018). *Combining Non-Probability and Probability Survey Samples through Mass Imputation*. Submitted, https://arxiv.org/abs/1812.10694

Kim, J. K., and J. N. K. Rao (2012). "Combining Data from Two Independent Surveys: A Model-Assisted Approach," *Biometrika*, 99, 85–100.

Kweon, S., Y. Kim, M-J Jang, Y. Kim, K. Kim, S. Choi, C. Chun, Y.-H. Khang, and K. Oh (2014). "Data Resource Profile: The Korea National Health and Nutrition Examination Survey (KNHANES)," *International Journal of Epidemiology*, 43, 69–77.

Lohr, S. L., and T. E. Raghunathan (2017). "Combining Survey Data with Other Data Sources," *Statistical Science*, 32, 293–312.

Nelder, J. A., and R. J. Baker (1972). *Generalized Linear Models*, New York: Wiley.

Rivers, D. (2007). "Sampling for Web Surveys," Proceedings of the Survey Research Methods Section of the American Statistical Association.

Ruppert, D., M. P. Wand, and R. J. Carroll (2009). "Semiparametric Regression during 2003–2007," *Electronic Journal of Statistics*, 3, 1193–1256.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*, Hoboken, NJ: Wiley.

Wang, D., and S. X. Chen (2009). "Empirical Likelihood for Estimating Equations with Missing Values," *The Annals of Statistics*, 37, 490–517.

Wang, L., B. I. Graubard, H. A. Katki, and Y. Li (2020). "Improving External Validity of Epidemiologic Cohort Analyses: A Kernel Weighting Approach," *Journal of the Royal Statistical Society Series A*, 183, 1293–1311.

Wang, Z., J. K. Kim, and S. Yang (2018). "An Approximate Bayesian Inference under Informative Sampling," *Biometrika*, 105, 91–102.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*, Chapman and Hall/CRC.

Wood, S. (2019). mgcv r package version 1.8-29.

Yang, S., and J. K. Kim (2018). "Integration of Survey Data and Big Observational Data for Finite Population Inference Using Mass Imputation," *arXiv Preprint arXiv*, 1807.02817.

Yang, S., J. K. Kim, and R. Song (2020). "Doubly Robust Inference When Combining Probability and Non-Probability Samples with High Dimensional Data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82, 445–465.