ORIGINAL ARTICLE

# Robust quasi-randomisation-based estimation with ensemble learning for missing data

Danhyang Lee[1] | Li-Chun Zhang[2] | Sixia Chen[3]

[1]Department of Information Systems, Statistics and Management Science, University of Alabama, Tuscaloosa, AL, 35487, USA

[2]Department of Social Statistics and Demography, University of Southampton, Southampton, UK

[3]Department of Biostatistics and Epidemiology, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma, 73104, USA

**Correspondence**
Danhyang Lee, Department of Information Systems, Statistics and Management Science, University of Alabama, Tuscaloosa, AL, 35487, USA
Email: dlee84@cba.ua.edu

**Funding information**
National Institute on Minority Health and Health Disparities at National Institutes of Health, Grand/Award Number: 1R21MD014658-01A1; Oklahoma Shared Clinical and Translational Resources with an Institutional Development Award (IDeA) from National Institute of General Medical Sciences, Grant/Award Number: U54GM104938

Missing data analysis requires assumptions about an outcome model or a response probability model to adjust for potential bias due to nonresponse. Doubly robust (DR) estimators are consistent if at least one of the models is correctly specified. Multiply robust (MR) estimators extend DR estimators by allowing for multiple models for both the outcome and/or response probability models, and are consistent if any of the multiple models is correctly specified. We propose a robust quasi-randomisation-based model approach to bring more protection against model misspecification than the existing DR and MR estimators, where any multiple semiparametric, nonparametric or machine learning models can be used for the outcome variable. The proposed estimator achieves unbiasedness by using a subsampling Rao-Blackwell method, given cell-homogenous response, but regardless of any working models for the outcome. An unbiased variance estimation formula is proposed, which does not use any replicate jackknife or bootstrap methods. Simulation study shows that our proposed method outperforms the existing multiply robust estimators.

KEYWORDS
item nonresponse, missing at random, cell mean model,

**Abbreviations:** DR, doubly robust; MR, multiply robust; MAR, missing at random; OR, outcome regression; RP, response probability; RB, Rao-Blackwell; SRB-MA, subsampling Rao-Blackwell model-assisted; SRB-MMA, subsampling Rao-Blackwell multiple model-assisted.

Rao-Blackwell method, variance estimation

## 1 | INTRODUCTION

Missing data can bring critical challenges in making valid inferences. It is well known that a direct application of statistical methods to complete cases without appropriate treatments of nonresponse could lead to significant biases, as the respondents often systematically differ from the non-respondents (Kim and Shao, 2021; Little and Rubin, 2019). To remove or reduce the biases due to nonresponse, model-based approach has been commonly used, where either an outcome (imputation) model for a study variable of interest or a response (probability) model to mimic the unknown response mechanism is assumed to be true. However, this approach is vulnerable to model misspecification.

An estimator is said to be doubly robust in the relevant literature if it remains asymptotically unbiased and consistent if either the outcome model or response probability model is correctly specified. Since DR estimators have double protection on asymptotic estimation consistency against model misspecifications, it has been widely used in missing data analysis. Kott (1994), Kott (2006), Kim and Park (2006), Haziza and Rao (2006), Kott and Chang (2010), Haziza et al. (2014), and Kim and Haziza (2014) discussed the DR procedures in the survey sampling context.

Those existing DR estimators, however, may fail to achieve consistent estimation in many practical studies. It allows only a single model for the study variable and a single model for the response probability. With an unknown true data-generating process, it is still risky to assume that one of the two models is correctly specified. This has motivated the development of a multiply robust estimator that Han and Wang (2013) first introduced, where multiple models for the outcome and response probability are considered in estimation. A multiply robust estimator is consistent if any one of those models is correctly specified, thus, can bring more protection against model misspecification than the DR procedures. For example, multiple models may be fitted in practice, each involving different subsets of covariates and possibly different link functions. Such models increase the likelihood of correct specification (Han and Wang, 2013; Chen and Haziza, 2017). Han and Wang (2013) used an empirical likelihood approach to develop a multiply robust point estimator and Han (2014) considered the case of regression analysis as an extension of Han and Wang (2013). Chen and Haziza (2017) developed multiply robust procedure in a finite population setting. See also Chen and Haziza (2021) for a review.

While Han and Wang (2013) and Chen and Haziza (2017) focus on expanding the pool of candidate parametric models with different sets of variables and/or different link functions, we aim to propose a new class of robust estimators for better performance by relaxing parametric model assumptions for the outcome variable so that any semiparametric, nonparametric, or machine learning models can be used as working models as well. Especially, many machine learning techniques can potentially be powerful assisting models, so our approach uses them as our working outcome models. In our procedure, multiple working models for the outcome variable are learned from a random subsample of the respondents, and their prediction errors unexplained by the outcome models are observed from the hold-out subsample of the respondents and projected to the non-respondents under the cell mean response probability model, which lead to multiple robust estimators. We define our final proposed estimator as a weighted average of those multiple estimators, where the weights are determined in a data-driven approach by using a subsampling Rao-Blackwell, so that the variance of the prediction errors, thus the variance of the estimator, can be minimized.

Our proposed approach extends some related ideas in machine learning and model-assisted estimation in the absence of missing data. Our use of multiple outcome models can be characterised as weighting-based ensemble learning (e.g. Zhou (2012)). It is similar in spirit to the super learner proposed by Van der Laan et al. (2007), which aims to improve prediction by creating a weighted combination of many candidate outcome learners, but in a different manner to ours. Whereas our use of the cell response model is a robust extension of the randomisation-based approach of unbiased statistical learning proposed by Sanguiao-Sande and Zhang (2021), which applies a single assisting outcome model to the complete sample observations by the subsampling Rao-Blackwell method.

We show that the proposed estimator achieves unbiasedness by using a subsampling Rao-Blackwell method, regardless of any working outcome models. Variance estimation for estimators relying on any nonparametric or machine learning models can be a challenging problem due to the complexity. We develop an unbiased variance estimation formula for our proposed estimator, which readily accommodates such models and does not use any replicate jackknife or bootstrap methods. We also note that although Han and Wang (2013) and Chen and Haziza (2017) develop their multiply robust estimators and related theoretical properties using parametric models, their procedures are able to include nonparametric or machine learning models via calibration constraints. We compare one of them to our proposed estimator in the survey sampling context, presented in Section 5.

The remainder of the paper proceeds as follows. In Section 2, a basic setup is introduced. In Section 3, we develop our proposed estimator in the case of a single outcome model and its variance estimator, and demonstrate their unbiasedness. We extend the proposed approach to allow for multiple outcome models in Section 4. A simulation study is given in Section 5 and concluding remarks are made in Section 6.

## 2 | BASIC SETUP

Consider a finite population of $N$ elements identified by a set of indices $U = \{1, \dots, N\}$, where $N$ is known. Let $y_i$ and $x_i$ be a study variable of interest and the vector of covariates associated with $y_i$ for each unit $i$, respectively. Let $s$ denote the set of indices for the elements in a sample selected by a probability sampling $p(s)$, where $\sum_s p(s) = 1$ over all possible samples from $U$. We assume that $x_i$ is always observed but $y_i$ is subject to missingness. Let the population quantity of interest be $\theta_N = g(y_1, \dots, y_N)$, and let $\hat{\theta}$ be a linear estimator of $\theta_N$ under complete response. For example, if we define $\hat{\theta}$ as follows,

$$\hat{\theta} = \sum_{i \in s} w_i y_i, \tag{1}$$

where $w_i = \text{pr}(i \in s)^{-1}$ is the inverse of the first-order inclusion probability of unit $i$, it is a design-unbiased estimator of the population total $\theta_N = \sum_{i=1}^N y_i$. Given the existence of missing data, we define $\delta_i$ as a response indicator, i.e., $\delta_i = 1$ if $y_i$ is observed, and $\delta_i = 0$ otherwise. We assume that the response mechanism is missing at random (MAR) in the sense of Rubin (1976) as follows:

$$\text{pr}(\delta_i = 1 \mid x_i, y_i) = \text{pr}(\delta_i = 1 \mid x_i).$$

An imputation-based estimator of $\theta_N$ is given by

$$\hat{\theta}_I = \sum_{i \in s} w_i \left\{ \delta_i y_i + (1 - \delta_i) \hat{y}_i^* \right\}, \tag{2}$$

where $\hat{y}_i^*$ denotes the imputed value used to replace the missing value $y_i$ and it can be constructed from an assumed outcome regression (OR) model for the study variable $y$ as follows:

$$E\{Y \mid x_i, \delta_i = 0\} = m(x_i; \beta), \tag{3}$$

where $m(x_i;\beta)$ is any pre-specified function of $\beta$, and $\beta$ is a vector of unknown parameters. Under MAR, the imputed estimator can replace the missing values by $y_i^* = m(x_i;\hat{\beta})$, where $\hat{\beta}$ is a consistent estimator of the true parameter $\beta$ under the OR model (3). If the model (3) is misspecified, the imputed estimator with $y_i^* = m(x_i;\hat{\beta})$ is biased.

For a more robust approach, we can adopt in addition a response probability (RP) model for $\Pr(\delta_i = 1 \mid x_i)$ as follows:

$$\mathrm{pr}(\delta_i = 1 \mid x_i) = p(x_i;\alpha), \tag{4}$$

for some $\alpha$, where $p(x_i;\alpha)$ is any pre-specified function of $\alpha$ and $\alpha$ is a vector of unknown parameters. Then, a class of the estimators can be given in the form of

$$\hat{\theta}_{dr}(\hat{\beta},\hat{\alpha}) = \sum_{i\in s} w_i \left[ m(x_i;\hat{\beta}) + \frac{\delta_i}{p(x_i;\hat{\alpha})}\{y_i - m(x_i;\hat{\beta})\}\right], \tag{5}$$

where $\hat{\beta}$ is a consistent estimator for $\beta$ under the OR model and $\hat{\alpha}$ is consistent for $\alpha$ under the RP model. Note that the estimator (5) is referred to as a doubly robust estimator in the sense that it can be consistent if either one of the two models (3) and (4) is correctly specified, e.g. by examining

$$\hat{\theta}_{dr}(\hat{\beta},\hat{\alpha}) - \hat{\theta} = \sum_{i\in s_r} w_i \frac{y_i - m(x_i;\hat{\beta})}{p(x_i;\hat{\alpha})} - \sum_{i\in s} w_i\{y_i - m(x_i;\hat{\beta})\}.$$

$\hat{\theta}_{DR}(\hat{\beta},\hat{\alpha})$ is a class of estimators, of which properties are determined by how to choose $(\hat{\beta},\hat{\alpha})$. Scharfstein et al. (1999) and Haziza and Rao (2006) used maximum likelihood approach to estimate $\alpha$ and then estimate $\beta$ by using ordinary or iteratively reweighted least square methods. Cao et al. (2009) used the optimal score equation based on influence function theory and Kim and Haziza (2014) used the same estimating equation for $\beta$ as in Scharfstein et al. (1999) and Haziza and Rao (2006), but proposed to use a calibration condition regarding $\partial m(x_i;\beta)/\partial\beta$ to choose $\alpha$, instead of using the maximum likelihood approach. Kim and Haziza (2014) showed that their proposed DR estimator has better efficiency than the other DR estimators proposed by Cao et al. (2009), Haziza and Rao (2006), and Tan (2006).

As Chen and Haziza (2017) discussed, any DR estimators still require that either model is correctly specified, which is not always desirable in practice. They proposed a multiply robust imputation procedure which allows for multiple OR models and multiple RP models with different subset of covariates and different link functions. Their proposed estimator is consistent if one of the OR models is true or one of the RP models is true. It is more robust than doubly robust estimators by increasing the likelihood of having a true model for either outcome or response probability in the pool of candidate parametric models. However, they are still not free of failure due to model misspecification, given that the true data generating process is always unknown.

Rather than increasing the number of candidate parametric models for the outcome and the response probability with the assumption that one of them is correctly specified, we can gain more robustness by relaxing the parametric model assumptions, while achieving unbiased variance estimation as well as point estimation, as described in the following section.

## 3 | PROPOSED METHOD

We propose a different approach to develop a new class of estimators that is more robust than (5).

Let $s = s_r \cup s_m$ be the bipartition of sample $s$, where $s_r$ is the set of indices with complete response and $s_m$ is the set of indices with nonresponse. Initially, we randomly split $s_r$ into training set $s_1$ and test set $s_2$ such that $s_r = s_1 \cup s_2$, where $s_1$ is selected from $s_r$ by simple random sampling, denoted by $q(s_1 \mid s_r, s)$. In the first phase, we fit a suitable working model to learn $y$ based on the training set $s_1$; in the second phase, we observe the error of the first-phase model in the test set $s_2$.

One can use any suitable working model for $y$, which can be any parametric, semiparametric, nonparametric or machine learning models such as regression tree, random forest, and any other learning models. Let $\mu(x_i; s_1)$ denote the predicted value of the study variable $y_i$ for unit $i$ with features $x_i$, which is trained on $\{(x_i, y_i) : i \in s_1\}$. For any $j \notin s_1$, we define the error of prediction by $\mu(x; s_1)$ as

$$e_j = y_j - \mu(x_j; s_1).$$

Then, we can re-express $\hat{\theta}$ in (1) as

$$\hat{\theta} = \sum_{i \in s} w_i y_i = \sum_{i \in s_1} w_i y_i + \sum_{i \in s_1^c} w_i \{\mu(x_i; s_1) + e_i\},$$

where the prediction errors $e_i$ are observed in $s_2$ but missing for $i \in s_m$ since $y_i$ is not observed in $s_m$. We now estimate the total of the prediction errors in $s_1^c$ by using a response probability model.

We assume a cell mean model for the response probability rather than a parametric model, as follows. Under MAR, we assume that the population $U$ is partitioned into $G$ cells, i.e., $U = U_1 \cup U_2 \cup \cdots \cup U_G$ such that

$$\mathrm{pr}(\delta_i = 1) = p_g, \quad \text{if } i \in U_g, \tag{6}$$

for $g = 1, \ldots, G$. The partition can be constructed by the quantiles of $x$ Im et al. (2018). This has similar effects to forming the response cells with the help of a fitted parametric model (4), but is more robust than adopting the parametric RP model directly, as described in Haziza and Beaumont (2017). Unlike Kim and Fuller (2004), who assume the cell mean model for the study variable $y$, we assume the cell mean model for the response indicator $\delta$, so that we can treat $\mathcal{F}_N = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ as fixed constants when it comes to variance estimation.

*Remark 1.* For the cell formation (6), one can also use any type of RP model, including a machine learning model such as classification tree and random forest, which are often more robust than parametric model approaches, as demonstrated via numerical studies in Section 5.

Now, we apply the doubly robust idea to the complement of $s_1$, i.e., $s_1^c = s_2 \cup s_m$, conditional on $s_1$ to construct an unbiased estimator. We first define the induced probability of subsampling $s_1$ from $s$ given by

$$p_1(s_1 \mid s) = \sum_{s_r : s_1 \subset s_r} q(s_1 \mid s_r, s) p(s_r \mid s), \tag{7}$$

Let $s_{1g}, s_{2g}, s_{rg}, s_{mg}$ and $s_g$ be the corresponding subsamples of units from $U_g$, whose sizes are $n_{1g}, n_{2g}, n_{rg}, n_{mg}$ and

$n_g$. The conditional response probability for $i \in s \setminus s_1$ given $s_1$ and $s$ is given by

$$p_{2g} = \text{pr}(i \in s_{2g} \mid s_1, s) = n_{2g}/(n_g - n_{1g}), \tag{8}$$

for $g = 1, \ldots, G$. Then, we define an estimator of $\theta_N$ as

$$\begin{aligned} \hat{\theta}^{(1)} &= \sum_{i \in s_1} w_i y_i + \sum_{i \in s \cap s_1^c} w_i \mu(x_i; s_1) + \sum_{g=1}^{G} \sum_{i \in s_g \cap s_{1g}^c} w_i \frac{\delta_i}{p_{2g}} \{y_i - \mu(x_i; s_1)\}, \\ &= \sum_{i \in s_1} w_i y_i + \sum_{i \in s \cap s_1^c} w_i y_i^* \end{aligned} \tag{9}$$

where $y_i^* = \mu(x_i; s_1) + \sum_{g=1}^{G} I(i \in s_g)\delta_i/p_{2g}\{y_i - \mu(x_i; s_1)\}$ is the imputed value of $y_i$. Here, $e_i = y_i - \mu(x_i; s_1)$ is observed if $i \in s_{2g}$.

The difference of (10) to the full-sample estimator $\hat{\theta}$ can be given as

$$\hat{\theta}^{(1)} - \hat{\theta} = \sum_{g=1}^{G} \left\{ \sum_{i \in s_{2g}} w_i \frac{e_i}{p_{2g}} - \sum_{i \in s_{2g} \cup s_{mg}} w_i e_i \right\}$$

Clearly, if $\mu(x_i; s_1)$ corresponds to the true outcome model, then $\hat{\theta}^{(1)} - \hat{\theta}$ will be approximately zero. Next, although $\mu(x_i; s_1)$ may be misspecified to a greater or lesser extent generally, the conditional expectation of $\hat{\theta}^{(1)} - \hat{\theta}$ given $s_1$ and $\mu(\cdot; s_1)$ is still zero, if $p_{2g}$ is the inclusion probability in $s_{2g}$ induced by the cell-response model (6) and the subsequent random split of $(s_1, s_2)$.

**Lemma 1** *Assume that the cell response model (6) holds, where the partition of $U$ ($U = U_1 \cup U_2 \cup \cdots \cup U_G$) is fixed and known. Then, regardless the choice of $\mu$ and the given sampling design, we have*

$$E(\hat{\theta}^{(1)}|s) = E_r\{E_q(\hat{\theta}^{(1)} \mid s_r, s) \mid s\} = E_1\{E_2(\hat{\theta}^{(1)} \mid s_1, s) \mid s\} = \sum_{i \in s} w_i y_i,$$

*where $w_i = \pi_i^{-1}$, and $E_r(\cdot \mid s)$ and $E_q(\cdot \mid s, s_r)$ denote the expectations with respect to the response probability and subsampling distributions, respectively. $E_1(\cdot \mid s)$ and $E_2(\cdot \mid s_1, s)$ are the expectations over the induced probability of sampling $s_1$ from $s$ and corresponding conditional response probability given $s_1$ and $s$, respectively.*

By virtue of Lemma 1 (proof in Appendix 1) one can safely adopt any *assisting* outcome model for $y$, although a better outcome model could lead to a small variance of the resulting estimator than a worse model, given the cell response model. However, $\hat{\theta}^{(1)}$ is only based on one random split of $s_r = s_1 \cup s_2$, which leads to additional variance due to learning from $s_1$ instead of $s_r$. As proposed by Sanguiao-Sande and Zhang (2021), we can reduce the variance of $\hat{\theta}^{(1)}$ by applying the Monte Carlo Rao-Blackwell (RB) method. Then, our proposed estimator is given by

$$\hat{\theta}_{SRB} = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}^{(k)} \tag{10}$$

where $\hat{\theta}^{(k)}$ is the estimator (10) calculated from the $k$-th Monte Carlo subsamples, $(s_1^{(k)}, s_2^{(k)})$ such that $s_r = s_1^{(k)} \cup s_2^{(k)}$, for $k = 1, \ldots, K$. It converges to $E_q(\hat{\theta}_{DR}^{(1)} \mid s, s_r)$, where the expectation is evaluated with respect to random subsampling of $s_1$ from $s_r$, i.e., $q(s_1 \mid s_r, s)$.

Theorem 1 below gives the properties of the proposed *subsampling Rao-Blackwell (SRB) quasi-randomisation-based* estimator $\hat{\theta}_{SRB}$ under the joint probability distribution induced by the sampling design, random subsampling, and the cell mean response probability model. See Appendix 1 for the proof.

**Theorem 1** *Assume that the cell response model (6) holds, where the partition of $U$ is fixed and known. Then, regardless the choice of $\mu$ and the given sampling design, the estimator $\hat{\theta}_{SRB}$ is unbiased for the finite population total $\theta_N$, i.e., $E(\hat{\theta}_{SRB}) = \theta_N$, and its variance is*

$$\begin{aligned}\mathrm{var}(\hat{\theta}_{SRB}) \quad &= \quad \mathrm{var}\left(\sum_{i \in s} w_i y_i\right) + E\{\mathrm{var}_2(\hat{\theta}^{(1)} \mid s_1, s)\} - E\{\mathrm{var}_q(\hat{\theta}^{(1)} \mid s, s_r)\} \\ &\quad\quad + E\{\mathrm{var}_q(\hat{\theta}_{SRB} \mid s, s_r)\},\end{aligned}$$

*where*

$$\mathrm{var}(\hat{\theta}^{(1)} \mid s_1, s) = \sum_{g=1}^{G} \frac{(n_g - n_{1g})^2}{n_{2g}} \left(1 - \frac{n_{2g}}{n_g - n_{1g}}\right) \frac{1}{n_g - n_{1g} - 1} \sum_{j \in (s \setminus s_1)_g} (w_j e_j - \bar{e}_{w,g})^2,$$

*and $\bar{e}_{w,g} = \sum_{j \in (s \setminus s_1)_g} w_j e_j / (n_g - n_{1g})$.*

Note that the first two terms account for the variance due to sampling and the variance due to the response and imputation, respectively. The third term, $E\{\mathrm{var}(\hat{\theta}^{(1)} \mid s)\}$, measures the variance reduction by $E(\hat{\theta}^{(1)} \mid s)$ rather than $\hat{\theta}^{(1)}$, and the last term is added due to the Monte Carlo RB method instead of exact RB, $E(\hat{\theta}^{(1)} \mid s)$.

Theorem 2 gives an unbiased estimator of $\mathrm{var}(\hat{\theta}_{SRB})$, denoted by $\hat{V}_{SRB}$. See Appendix 1 for the proof.

**Theorem 2** *Let $\pi_{ij}$ be the second-order sample inclusion probability, and $\pi_i > 0$ for all $i \in U$, and $\pi_{ij} > 0$ for all $i, j \in U$. We define an estimator of $\mathrm{var}(\hat{\theta}_{SRB})$ as*

$$\hat{V}_{SRB} \quad = \quad \hat{V}^1 + \hat{V}^2 - \hat{V}^3 + \hat{V}^4$$

*where*

$$\begin{aligned}\hat{V}^1 \quad &= \quad \sum_{i \in s_r} \sum_{j \in s_r} \frac{1}{\hat{p}_{ij}} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}, \\ \hat{V}^2 \quad &= \quad \frac{1}{K} \sum_{k=1}^{K} \sum_{g=1}^{G} \frac{(n_g - n_{1g}^{(k)})^2}{n_{2g}^{(k)}} \left(1 - \frac{n_{2g}^{(k)}}{n_g - n_{1g}^{(k)}}\right) \frac{1}{(n_{2g}^{(k)} - 1)} \sum_{j \in s_{2g}^{(k)}} (w_j e_j^{(k)} - \bar{e}_{w,g}^{(k)})^2, \\ \hat{V}^3 \quad &= \quad \frac{1}{K-1} \sum_{k=1}^{K} (\hat{\theta}^{(k)} - \hat{\theta}_{SRB})^2, \\ \hat{V}^4 \quad &= \quad \frac{1}{K(K-1)} \sum_{k=1}^{K} (\hat{\theta}^{(k)} - \hat{\theta}_{SRB})^2,\end{aligned}$$

*where $\hat{p}_{ij} = \widehat{\mathrm{pr}}(\delta_i \delta_j = 1 \mid x_i, x_j)$ under the cell response model (6), and $\bar{e}_{w,g}^{(k)} = \sum_{j \in s_{2g}^{(k)}} w_j e_j^{(k)} / n_{2g}^{(k)}$. The estimator $\hat{V}$ is unbiased for* $\mathrm{var}(\hat{\theta}_{SRB})$.

## 4 | EXTENSION TO MULTIPLY ROBUST ESTIMATION

The SRB estimator (10) relies on only one working model for $y$. Instead of choosing a single outcome model, we propose another SRB-based estimator that takes multiple outcome models into account as in (11) below, which is referred to as *subsampling Rao-Blackwell ensemble learning-assisted estimator (SRB-EL)*:

$$\hat{\theta}_{SRBEL} = \frac{1}{K}\sum_{k=1}^{K}\sum_{m=1}^{M} a_m^{(k)}\hat{\theta}_m^{(k)}, \tag{11}$$

where $a_m^{(k)} \in \{0, 1\}$ is an indicator of selecting model $\mu_m(x; \cdot)$, such that $\sum_{m=1}^{M} a_m^{(k)} = 1$, and $\hat{\theta}_m^{(k)}$ is the SRB estimator based on model $m$ from the $k$-th Monte Carlo subsamples ($k = 1, \dots, K$).

In this study, we define $a_m^{(k)} = 1$ if model $\mu_m(x; s_1^{(k)})$ has the lowest prediction errors in $s_2^{(k)}$ in terms of $\hat{V}_2^{(m,k)}$ given by

$$\hat{V}_2^{(m,k)} = \sum_{g=1}^{G}\frac{(n_g - n_{1g}^{(k)})^2}{n_{2g}^{(k)}}\left(1 - \frac{n_{2g}^{(k)}}{n_g - n_{1g}^{(k)}}\right)\frac{1}{(n_{2g}^{(k)} - 1)}\sum_{j\in s_{2g}^{(k)}}(w_j e_j^{(m,k)} - \bar{e}_{w,g}^{(m,k)})^2, \tag{12}$$

where $e_j^{(m,k)} = y_j - \mu_m(x_j; s_1^{(k)})$ and $\bar{e}_{w,g}^{(m,k)} = \sum_{j\in s_{2g}^{(k)}} w_j e_j^{(k)}/n_{2g}^{(k)}$. Recall that $\hat{V}_2^{(m,k)}$ $(k = 1, \dots, K)$ sums up to $\hat{V}^2$ for model $m$ in Theorem 2 that measures the variance due to imputation by model $m$ as well as the response probability. Thus, the proposed SRB-EL estimator gives more weight to a model that has a smaller imputation variance. As $K$ goes to infinity, $\hat{\theta}_{SRBEL}$ will tend to $\hat{\theta}^*_{SRBEL} = \sum_{m=1}^{M} p_m\hat{\theta}^*_m$, where $p_m$ is the probability of choosing model $m$ such that $\sum_{m=1}^{M} p_m = 1$, and $\hat{\theta}^*_m = E_q(\hat{\theta}_m^{(1)} \mid s, s_R)$.

**Proposition 1** *Assume that the cell mean response model (6) holds, where the partition of $U$ is fixed and known. For any multiple $K$ working outcome models $\mu_m (m = 1, \dots, M)$ and the given sampling design, we have*

$$E(\hat{\theta}_{SRBEL}) = \theta_N$$

*and*

$$\begin{aligned}\mathrm{var}(\hat{\theta}_{SRBEL}) \quad = \quad & \mathrm{var}\left(\sum_{i\in s} w_i y_i\right) + E\left\{\mathrm{var}_2\left(\bar{\theta}_m^{(1)} \mid s_1, s, a\right)\right\} - E\left\{\mathrm{var}_q\left(\bar{\theta}_m^{(1)} \mid s, s_r\right)\right\} \\ & + E\left\{\mathrm{var}_q(\hat{\theta}_{SRBEL} \mid s, s_r)\right\},\end{aligned}$$

*where $\bar{\theta}_m^{(1)} = \sum_{m=1}^{M} a_m\hat{\theta}_m^{(1)}$, $a = (a_1, \dots, a_M)^T$ and*

$$\mathrm{var}_2(\bar{\theta}_m^{(1)} \mid s_1, s, a) = \sum_{g=1}^{G}\frac{(n_g - n_{1g})^2}{n_{2g}}\left(1 - \frac{n_{2g}}{n_g - n_{1g}}\right)\frac{1}{n_g - n_{1g} - 1}\sum_{j\in(s\setminus s_1)_g}(w_j\bar{e}_j^{(m)} - \bar{e}_{w,g}^{*(m)})^2,$$

*for $m = 1, \dots, M$. Here, $\bar{e}_{w,g}^{*(m)} = \sum_{j\in(s\setminus s_1)_g} w_j\bar{e}_j^{(m)}/(n_g - n_{1g})$, $\bar{e}_j^{(m)} = \sum_{m=1}^{M} a_m e_j^{(m)}$, and $e_j^{(m)} = y_j - \mu_m(x_j; s_1)$ is the prediction error by the $m$-th outcome model for unit $j \in s_1^c$.*

**Proposition 2** *Let $\pi_{ij}$ be the second-order sample inclusion probability, and $\pi_i > 0$ for all $i \in U$, and $\pi_{ij} > 0$ for all $i, j \in U$.*

*The proposed variance estimator given below is unbiased for* $\mathrm{var}(\hat{\theta}_{SRBMR})$*:*

$$\hat{V}_{SRBEL} = \hat{V}^1 + \hat{V}^2 - \hat{V}^3 + \hat{V}^4,$$

*where*

$$\begin{aligned}
\hat{V}^1 &= \sum_{i \in s_r} \sum_{j \in s_r} \frac{1}{\hat{p}_{ij}} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}, \\
\hat{V}^2 &= \frac{1}{K} \sum_{k=1}^{K} \sum_{g=1}^{G} \frac{(n_g - n_{1g}^{(k)})^2}{n_{2g}^{(k)}} \left(1 - \frac{n_{2g}^{(k)}}{n_g - n_{1g}^{(k)}}\right) \frac{1}{(n_{2g}^{(k)} - 1)} \sum_{j \in s_{2g}^{(k)}} (w_j \bar{e}_j^{(k)} - \bar{e}_{w,g}^{*(k)})^2, \\
\hat{V}^3 &= \frac{1}{K-1} \sum_{k=1}^{K} (\hat{\theta}_{EL}^{(k)} - \hat{\theta}_{SRBEL})^2, \\
\hat{V}^4 &= \frac{1}{K(K-1)} \sum_{k=1}^{K} (\hat{\theta}_{EL}^{(k)} - \hat{\theta}_{SRBEL})^2,
\end{aligned}$$

*where* $\hat{p}_{ij} = \widehat{\mathrm{pr}}(\delta_i \delta_j = 1 \mid x_i, x_j)$ *under the cell response model (6) with fixed and known imputation cells,* $\bar{e}_{w,g}^{*(k)} = \sum_{j \in s_{2g}^{(k)}} w_j \bar{e}_j^{(k)} / n_{2g}^{(k)}$*, and* $\bar{e}_j^{(k)} = \sum_{m=1}^{M} a_m^{(k)} e_j^{(m,k)}$.

*Remark 2.* The proposed SRB-EL estimator is based on a single cell RP model, which can actually be formed based on one or several RP models, but is more robust than applying the RP models directly. Moreover, the estimator adaptively assigns different weights to the multiple outcome models, according to their prediction errors that are observed in the hold-out test subsample, so that it is more robust than assuming any of them to be true. In this sense, the proposed estimator can be considered as multiply robust, or simply robust by the usage of the term in robust statistics. This differs to the multiply robustness defined in Han and Wang (2013), where the estimator is good if one of the OR or RP models is correct but without the assurance that the performance does not deteriorate considerably otherwise.

## 5 | SIMULATION STUDY

We conduct a simulation study to evaluate the performance of the proposed method using a similar simulation setup as Chen and Haziza (2017) which followed the setup of Kang and Schafer (2007).

In Scenario 1, we generate 1,000 finite populations of size $N = 10,000$ as follows. For unit $i (i = 1, \dots, N)$, a vector $x_i = (x_{1i}, x_{2i}, x_{3i})^T$ is randomly generated, where $x_{1i} - 1 \sim \mathrm{Poisson}(5)$ and $(x_{2i}, x_{3i})^T$ is generated from a standard multivariate normal distribution, and $y_i = 5 - 1.5x_{1i} + x_{3i} + \epsilon_i$ is generated accordingly, where $\epsilon_i$ is a standard normal random error associated with unit $i$.

From each finite population, we select a random sample of size $n = 800$ using random sampling with probability proportional to a size variable, $\psi_i = 0.5\chi_i + 1$, where $\chi_i$ is generated from a chi-square distribution with one degree of freedom. The inclusion probability is $\pi_i = n\psi_i / \sum_{j \in U} \psi_j$, for $i \in \{1, \dots, N\}$.

In each sample, missing $y$ is generated with probability $\mathrm{pr}(\delta_i = 0 \mid x_i) = \{1 + \exp(\alpha_0 + \alpha_1 x_{1i} + \alpha_1 x_{2i} + \alpha_3 x_{3i})\}^{-1}$, where we set $(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (-0.7, 0.2, -0.5, 0.5)$ which leads to the average response rate of 60%.

In addition, we consider the following transformations of the $x$-variables: $z_{1i} = \exp(x_{1i}/2)$, $z_{2i} = x_{2i} / \{1 + \log(x_{1i})\}$, and $z_{3i} = x_{2i} x_{3i}$. Letting $p(x; \alpha) = 1 - \{1 + \exp(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i})\}^{-1}$ and $m(x; \beta) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$

denote the correct response probability and OR models, we consider $p(z;\tilde{\alpha})$ and $m(z;\tilde{\beta})$ as incorrect models in this scenario.

From each realized incomplete samples, we compute the following estimators for $\theta_N = N^{-1}\sum_{i=1}^{N} y_i$.

1. $\bar{y}_{Full}$: As a gold benchmark, we use the full samples and compute $\hat{\theta} = N^{-1}\sum_{i\in s} w_i y_i$
2. $\bar{y}_{CH}(****)$: As one of the existing multiply robust estimators, we compute the estimator of Chen and Haziza (2017) which is developed in the survey sampling framework. The four digits in parentheses indicate which models are used for estimation. The first two digits correspond to $m(x;\beta)$ and $m(z;\tilde{\beta})$, and the last two digits correspond to $p(x;\alpha)$ and $p(z;\tilde{\alpha})$, respectively.
3. $\bar{y}_{CH:++}(****)$: Although Chen and Haziza (2017) focused on parametric models for the outcome and response probability, we note that their estimation procedure can incorporate any nonparametric or machine learning models into the calibration constraints, which is used for a further investigation. In addition to $\bar{y}_{CH}(****)$, we consider the following models in estimation:
   - $\bar{y}_{CH:CM}(****)$: The cell mean model (6) assumed for the response probability, where we use all the six explanatory variables $(x_1, x_2, x_3, z_1, z_2, z_3)$ and construct 9 cells based on the random forest classification for $p_i$. Using a larger number of cells does not bring any significant difference.
   - $\bar{y}_{CH:RF}(****)$: A random forest on $(x, y)$ used for the outcome variable.
   - $\bar{y}_{CH:RFCM}(****)$: Both the cell mean model (6) and random forest on $(x, y)$ are added to the calibration constraints.
6. $\bar{y}_{SRBEL}$: We compute our proposed estimator (11) with three outcome models, $m(x;\beta)$, $m(z;\tilde{\beta})$ and a random forest on $(x, y)$, and the response cells are constructed in the same way as in $\bar{y}_{CH}$. We apply the Monte Carlo RB method with $K = 50$ and a 50-50 random split between $s_1^{(k)}$ and $s_2^{(k)}$ in $s_r$ for $k = 1, \dots, K$.

In Scenario 2, we generate 1,000 finite populations of size $N = 10,000$ as in Scenario 1 except that we now use $y_i = 5 + 1.5x_{1i} - 2x_{2i}^2 + x_{3i} + \epsilon_i$, and $\mathrm{pr}(\delta_i = 0 | \ x_i) = 1 - \Phi(-1.5 + 0.2x_{1i} + 0.7x_{2i} + 0.25x_{3i})$, where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. Note that none of the six models, $p(z;\alpha)$, $p(z;\tilde{\alpha})$, the cell mean model, $m(x;\beta)$, $m(z;\tilde{\beta})$ and the random forest, are correctly specified.

Table 1 presents the Monte Carlo biases, standard errors, and root mean squared errors of the estimators of $\theta_N$ under both the scenarios. In Scenario 1 where $m(x;\beta)$ and $p(x;\alpha)$ are true models for $y$ and $\delta$, the multiply robust estimators based on the four parametric models except for $\bar{y}_{CH}(0101)$, show negligible biases, since at least one of the outcome and response probability models is correctly specified. $\bar{y}_{CH:CM}(1000)$ also works well since the true outcome model is used, while $\bar{y}_{CH:RF}(0010)$ shows a lower efficiency with a small increase in bias in spite of using the correct response probability model. Note that $\bar{y}_{CH:RFCM}(1100)$ is the most comparable estimator to our proposed estimator $\bar{y}_{CH:RFCM}$ in a sense that both use the three outcome models and one cell mean response model for estimation. We can see that our proposed estimator is more efficient than $\bar{y}_{CH:RFCM}$. Comparing between $\bar{y}_{CH:RFCM}(1111)$ and $\bar{y}_{CH}(1111)$, adding the random forest and cell mean model into the calibration constraints does not bring any improvement in both efficiency and bias reduction.

In Scenario 2, none of working OR or RP model is correctly specified. As expected, all of the MR estimators based on the four parametric models ($\bar{y}_{CH}$) have large biases. However, either using the cell mean response model or random forest for the outcome, as in $\bar{y}_{CH:CM}(1000)$, $\bar{y}_{CH:CM}(0100)$, $\bar{y}_{CH:RF}(0010)$, and $\bar{y}_{CH:RF}(0001)$, greatly reduced the biases observed from $\bar{y}_{CH}(****)$. Especially, $\bar{y}^*_{CH:RFCM}(1100)$ and $\bar{y}_{CH:RFCM}(1111)$ show negligible biases, but the proposed estimator is seen to achieve the best performance in both efficiency and bias.

All the estimators considered in the simulations can perform reasonably, provided one builds good OR and RP

**TABLE 1** Monte Carlo root mean squared errors (RMSE), standard errors and biases of the several estimators of $\theta_N$ based on 1,000 Monte Carlo samples

| Estimator | Scenario 1 | | | Scenario 2 | | |
|---|---|---|---|---|---|---|
| | RMSE | SE | Bias | RMSE | SE | Bias |
| $\bar{y}_{Full}$ | 0.787 | 0.787 | 0.028 | 0.480 | 0.480 | 0.020 |
| $\bar{y}_{CH}(1010)$ | 0.790 | 0.789 | 0.028 | 1.125 | 0.538 | -0.988 |
| $\bar{y}_{CH}(1001)$ | 0.789 | 0.789 | 0.028 | 1.070 | 0.472 | -0.960 |
| $\bar{y}_{CH}(0110)$ | 0.801 | 0.799 | 0.053 | 1.096 | 0.580 | -0.930 |
| $\bar{y}_{CH}(0101)$ | 0.783 | 0.717 | 0.315 | 1.334 | 0.912 | -0.974 |
| $\bar{y}_{CH}(1111)$ | 0.790 | 0.789 | 0.029 | 1.586 | 0.943 | -1.275 |
| $\bar{y}_{CH:CM}(1000)$ | 0.791 | 0.790 | 0.027 | 0.444 | 0.321 | -0.306 |
| $\bar{y}_{CH:CM}(0100)$ | 0.817 | 0.777 | 0.252 | 0.330 | 0.295 | -0.146 |
| $\bar{y}_{CH:RF}(0010)$ | 0.888 | 0.878 | 0.128 | 0.517 | 0.511 | -0.079 |
| $\bar{y}_{CH:RF}(0001)$ | 0.897 | 0.885 | 0.143 | 0.542 | 0.523 | -0.142 |
| $\bar{y}_{CH:RFCM}(0000)$ | 0.889 | 0.875 | 0.158 | 0.511 | 0.504 | -0.082 |
| $\bar{y}^*_{CH:RFCM}(1100)$ | 0.880 | 0.867 | 0.150 | 0.517 | 0.516 | -0.034 |
| $\bar{y}_{CH:RFCM}(1111)$ | 0.885 | 0.876 | 0.129 | 0.513 | 0.511 | -0.046 |
| $\bar{y}_{SRBEL}$ | 0.788 | 0.788 | 0.028 | 0.484 | 0.484 | -0.020 |

RMSE, Root mean squared error; SE, Standard error.

$\bar{y}^*_{CH:RFCM}(1100)$ is most comparable to $\bar{y}_{proposed}$, in a sense that the same candidate models (three outcome models and cell mean response model) are used.

models to start with, as seen from Scenario 1. The problem is that this may be difficult to achieve in practice, at least when restricting to the models most commonly applied in survey sampling, or when there are a large number of covariates to work with. Although the models used by the SRB-EL estimator are not fined-tuned manually, gains are readily obtained by allowing for more flexible models such as random forest that do not require too much fine-tuning to be reasonable, now that the SRB method allows one to move freely beyond familiar parametric models.

Finally, we have calculated the relative bias (RB) of the proposed variance estimator, $\hat{V}_{SRBEL}$, in Proposition 2, as follows.

$$RB(\%) = \frac{E_{MC}(\hat{V}_{SRBEL}) - \text{var}_{MC}(\bar{y}_{SRBEL})}{\text{var}_{MC}(\bar{y}_{SRBEL})} \times 100,$$

where $E_{MC}(\cdot)$ and $\text{var}_{MC}(\cdot)$ denote the Monte Carlo mean and variance, respectively. We also calculate the coverage rate of a normal confidence interval with our proposed variance estimator, $\bar{y}_{SRBEL} \pm z_{\alpha/2}\hat{V}^{1/2}_{SRBEL}$, where $z_{\alpha/2}$ denotes the upper $(1-\alpha/2)$ critical value for the standard normal distribution and $\alpha = 0.05$. As can be seen from Table 2, the proposed variance estimator performs as intended.

**TABLE 2** Relative bias (RB) and coverage rate (%) of the proposed variance estimator ($\hat{V}$) of the SRB-EL estimator with the random-forest cell formation under two scenarios based on 1,000 Monte Carlo samples

| Estimator | Scenario | VE | MC Var | RB(%) | CR(%) |
|---|---|---|---|---|---|
| $\bar{y}_{SRBEL}$ | 1 | 0.616 | 0.621 | -0.80 | 94.5 |
|  | 2 | 0.228 | 0.234 | -2.36 | 94.3 |

VE, Variance estimate; MC Var, Monte Carlo variance; RB, Relative Bias; CR, Coverage rate.

## 6 | CONCLUDING REMARKS

We propose a new class of robust quasi-randomisation-based estimators, where multiple working models for the outcome can be any semiparametric, nonparametric or machine learning models, while a cell mean model is used for the response probability, which as well can be formed with the help of one or several RP models. Multiple outcome models are learned from a random subsample of the respondents, and the prediction errors unexplained by the outcome models are observed from the hold-out subsample of the respondents and projected to the non-respondents under the cell mean response probability model. The resulting estimator is unbiased given cell-homogenous response, regardless of any working outcome models and misspecification of the predictors. The proposed algorithm of Monte Carlo RB is easy to implement. The unbiased variance estimation formula is provided, which does not require any replicate jackknife or bootstrap methods.

The theoretical properties of the proposed estimators and variance estimation formulas are derived for a fixed finite population. If we consider superpopulation inference, our proposed estimators can become doubly and multiply robust estimators, in the sense that if one of multiple outcome models and cell RP model is correctly specified, the proposed estimators would be consistent, with suitable regularity conditions for random vectors $(x, y)$ and for working outcome models $\mu_m(x; \cdot)$. In this case, variance estimation can be developed asymptotically. Establishing such doubly and multiply robustness from our proposed estimator and asymptotic variance estimation under a superpopulation model will be further studied, particularly when machine learning models are allowed.

This study suggests other future work related to the methodology. Finding optimal cell formation in the cell mean model is an area for future work. Although the cell mean model for the response probability works well in our limited simulations, it can be relaxed for further improvement. Ongoing research involves a general modeling for the response probability in multiply robust estimation based on statistical learning. Extension to high-dimensional and/or multivariate data are also interesting research topics.

## ACKNOWLEDGMENTS

Dr. Sixia Chen was partially supported by the National Institute on Minority Health and Health Disparities at National Institutes of Health (1R21MD014658-01A1) and 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 official views of the National Institutes of Health.

# APPENDIX A. Proofs

**Proof of Lemma 1.**

$$\begin{aligned} E(\hat{\theta}^{(1)} \mid s) &= E_r\{E_q(\hat{\theta}^{(1)} \mid s, s_r) \mid s\} = E_1\{E_2(\hat{\theta}^{(1)} \mid s, s_1) \mid s\} \\ &= E\left\{\sum_{i\in s_1} w_i y_i + \sum_{g=1}^{G} \sum_{i\in(s\setminus s_1)_g} w_i(\mu(x_i; s_1) + e_i)\middle| s\right\} \\ &= E\left\{\sum_{i\in s_1} w_i y_i + \sum_{g=1}^{G} \sum_{i\in(s\setminus s_1)_g} w_i y_i \middle| s\right\} \\ &= E\left\{\sum_{g=1}^{G} \sum_{i\in s_g} w_i y_i \middle| s\right\} = \sum_{i\in s} w_i y_i. \end{aligned}$$

**Proof of Theorem 1.** By using Lemma 1 and the IID construction of $\{\hat{\theta}^{(k)} : k = 1, \ldots, K\}$, we have

$$E(\hat{\theta}_{SRB}) = \frac{1}{K}\sum_{k=1}^{K} E\left\{E(\hat{\theta}^{(k)} \mid s)\right\} = \frac{1}{K}\sum_{k=1}^{K} E\left(\sum_{i\in s} w_i y_i\right) = \theta_N.$$

For the variance of $\hat{\theta}$, we have

$$\begin{aligned} \text{var}(\hat{\theta}_{SRB}) &= \text{var}\{E_q(\hat{\theta}_{SRB} \mid s, s_r)\} + E\{\text{var}_q(\hat{\theta}_{SRB} \mid s, s_r)\} \\ &= \text{var}(\hat{\theta}^*_{SRB}) + E\{\text{var}_q(\hat{\theta}_{SRB} \mid s, s_r)\} \\ &= \text{var}\left(\sum_{i\in s} w_i y_i\right) + E\{\text{var}(\hat{\theta}^*_{SRB} \mid s)\} + E\{\text{var}_q(\hat{\theta}_{SRB} \mid s, s_r)\}, \end{aligned}$$

where $\hat{\theta}^*_{SRB} = E_q(\hat{\theta}^{(1)} \mid s, s_r)$, and the third equality holds because $\text{var}(\hat{\theta}^*_{SRB}) = \text{var}\{E(\hat{\theta}^*_{SRB} \mid s)\} + E\{\text{var}(\hat{\theta}^*_{SRB} \mid s)\}$ and by Lemma 1.

Note that

$$\text{var}(\hat{\theta}^*_{SRB} \mid s) = \text{var}(\hat{\theta}^{(1)} \mid s) - E\{\text{var}_q(\hat{\theta}^{(1)} \mid s, s_r) \mid s\}$$

and

$$\text{var}(\hat{\theta}^{(1)} \mid s) = E\{\text{var}_2(\hat{\theta}^{(1)} \mid s, s_1) \mid s\} + V\{E_2(\hat{\theta}^{(1)} \mid s, s_1) \mid s\} = E\{\text{var}_2(\hat{\theta}^{(1)} \mid s, s_1) \mid s\},$$

therefore, the variance result in Theorem 1 follows.

**Proof of Theorem 2.**

$$E(\hat{V}^1) = E\{E(\hat{V}^1 \mid s_1, s)\} = E\left\{\sum_{i\in s}\sum_{j\in s} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}\right\} = \text{var}\left(\sum_{i\in s} w_i y_i\right),$$

where the second equality holds by the construction of $\hat{p}_{ij}$.

$$
\begin{aligned}
E(\hat{V}^2) &= E\{E(\hat{V}^2 \mid s_1, s)\} \\
&= E\left\{\sum_{g=1}^{G} \frac{(n_g - n_{1g})^2}{n_{2g}} \left(1 - \frac{n_{2g}}{n_g - n_{1g}}\right) \frac{1}{(n_g - n_{1g} - 1)} \sum_{j \in (s \setminus s_1)_g} (w_j e_j - \bar{e}_{w,g})^2\right\}, \\
&= E\{\mathrm{var}(\hat{\theta}^{(1)} \mid s_1, s)\},
\end{aligned}
$$

where the second equality holds by the construction of $\hat{p}_{2g}$, for $g = 1, \ldots, G$. We can show that $E(\hat{V}^3) = E\{\mathrm{var}_q(\hat{\theta}^{(1)} \mid s, s_r)\}$ and $E(\hat{V}^4) = E\{\mathrm{var}_q(\hat{\theta}_{SRB} \mid s, s_r)\}$ due to the IID construction of $\{\hat{\theta}^{(k)} : k = 1, \ldots, K\}$.

**Proof of Proposition 1.** By using the same argument as in the proof of Theorem 1, we have

$$
\mathrm{var}(\hat{\theta}_{SRBEL}) = \mathrm{var}\left(\sum_{i \in s} w_i y_i\right) + E\{\mathrm{var}(\hat{\theta}^*_{SRBEL} \mid s)\} + E\{\mathrm{var}_q(\hat{\theta}_{SRBEL} \mid s, s_r)\},
$$

where $\hat{\theta}^*_{SRBEL} = E_q(\sum_{m=1}^{M} a_m^{(1)} \hat{\theta}_m^{(1)} \mid s, s_r) = \sum_{m=1}^{M} p_m \hat{\theta}^*_m$, $\hat{\theta}^*_m = E_q(\hat{\theta}_m^{(1)} \mid s, s_r)$ and $p_m = \Pr(a_m = 1 \mid s, s_r)$. Since

$$
\mathrm{var}(\hat{\theta}^*_{SRBEL} \mid s) = \mathrm{var}\left(\sum_{m=1}^{M} a_m^{(1)} \hat{\theta}_m^{(1)} \mid s\right) - E\left\{\mathrm{var}_q\left(\sum_{m=1}^{M} a_m^{(1)} \hat{\theta}_m^{(1)} \mid s, s_r\right)\right\},
$$

and

$$
\begin{aligned}
\mathrm{var}\left(\sum_{m=1}^{M} a_m^{(1)} \hat{\theta}_m^{(1)} \mid s\right) &= E\left\{\mathrm{var}\left(\sum_{m=1}^{M} a_m^{(1)} \hat{\theta}_m^{(1)} \mid s, s_1, a\right) \mid s\right\} + \mathrm{var}\left\{E\left(\sum_{m=1}^{M} a_m^{(1)} \hat{\theta}_m^{(1)} \mid s, s_1, a\right) \mid s\right) \\
&= E\left\{\mathrm{var}\left(\sum_{m=1}^{M} a_m^{(1)} \hat{\theta}_m^{(1)} \mid s, s_1, a\right) \mid s\right\},
\end{aligned}
$$

where $a = (a_1^{(1)}, \ldots, a_M^{(1)})^T$, the variance result in the proposition follows.

## references

Cao, W., Tsiatis, A. 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.

Chen, S. and Haziza, D. (2017) Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika*, **104**, 439–453.

— (2021) A review of multiply robust estimation with missing data. *Modern Statistical Methods for Health Research*, 103–118.

Han, P. (2014) Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association*, **109**, 1159–1173.

Han, P. and Wang, L. (2013) Estimation with missing data: beyond double robustness. *Biometrika*, **100**, 417–430.

Haziza, D. and Beaumont, J.-F. (2017) Construction of weights in surveys: A review. *Statistical Science*, **32**, 206–226.

Haziza, D., Nambeu, C.-O. and Chauvet, G. (2014) Doubly robust imputation procedures for finite population means in the presence of a large number of zeros. *Canadian Journal of Statistics*, **42**, 650–669.

Haziza, D. and Rao, J. N. (2006) A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology*, **32**, 53.

Im, J., Cho, I. H. and Kim, J.-K. (2018) Fhdi: An r package for fractional hot deck imputation. *R J.*, **10**, 140.

Kang, J. D. and Schafer, J. L. (2007) Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 523–539.

Kim, J. K. and Fuller, W. (2004) Fractional hot deck imputation. *Biometrika*, **91**, 559–578.

Kim, J. K. and Haziza, D. (2014) Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, **24**, 375–394.

Kim, J. K. and Park, H. (2006) Imputation using response probability. *Canadian Journal of Statistics*, **34**, 171–182.

Kim, J. K. and Shao, J. (2021) *Statistical methods for handling incomplete data*. Chapman and Hall/CRC.

Kott, P. S. (1994) A note on handling nonresponse in sample surveys. *Journal of the American Statistical Association*, **89**, 693–696.

— (2006) Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, **32**, 133.

Kott, P. S. and Chang, T. (2010) Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, **105**, 1265–1275.

Van der Laan, M. J., Polley, E. C. and Hubbard, A. E. (2007) Super learner. *Statistical applications in genetics and molecular biology*, **6**.

Little, R. J. and Rubin, D. B. (2019) *Statistical analysis with missing data*, vol. 793. John Wiley & Sons.

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

Sanguiao-Sande, L. and Zhang, L.-C. (2021) Design-unbiased statistical learning in survey sampling. *Sankhya A: The Indian Journal of Statistics*, **83**, 714–744.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, **94**, 1096–1120.

Tan, Z. (2006) A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, **101**, 1619–1637.

Zhou, Z. H. (2012) *Ensemble methods: foundations and algorithms*. CRC press.