Stat Comput
DOI 10.1007/s11222-010-9223-y

A comparison study of nonparametric imputation methods

Jianhui Ning · Philip E. Cheng

Received: 24 November 2009 / Accepted: 14 December 2010

Abstract Consider estimation of a population mean of a response variable when the observations are missing at random with respect to the covariate. Two common approaches to imputing the missing values are the nonparametric regression weighting method and the Horvitz-Thompson (HT) inverse weighting approach. The regression approach includes the kernel regression imputation and the nearest neighbor imputation. The HT approach, employing inverse kernel-estimated weights, includes the basic estimator, the ratio estimator and the estimator using inverse kernel-weighted residuals. Asymptotic normality of the nearest neighbor imputation estimators is derived and compared to kernel regression imputation estimator under standard regularity conditions of the regression function and the missing pattern function. A comprehensive simulation study shows that the basic HT estimator is most sensitive to discontinuity in the missing data patterns, and the nearest neighbors estimators can be insensitive to missing data patterns unbalanced with respect to the distribution of the covariate. Empirical studies show that the nearest neighbor imputation method is most effective among these imputation methods for estimating a finite population mean and for classifying the species of the iris flower data.

Keywords Classification · Inverse weighting · Kernel regression · Missing at random · Nearest neighbor imputation

J. Ning
Department of Mathematics and Statistics, Central China Normal University, Wuhan, China

P.E. Cheng (✉)
Institute of Statistical Science, Academia Sinica, Taipei, Taiwan
e-mail: pcheng@stat.sinica.edu.tw

1 Introduction

Proportional weighting and the Horvitz-Thompson (HT) inverse weighting (Horvitz and Thompson 1952) for estimating a population parameter are commonly used in the analysis of stratified sampling (Cochran 1977). Without assuming a parametric model, a nonparametric regression approach to estimating a population mean can be fairly efficient when the underlying joint distribution satisfies certain regularity conditions. The validity of these conditions may not be easily tested when the observed data are incomplete or partly missing. While nonparametric inference can hardly be examined for cases without regularity conditions, it is nevertheless useful to investigate the difference in computational performance between the nonparametric regression estimation and the HT estimation. This aims to compare the methods of estimation for missing data beyond the usual regularity conditions on the underlying distribution.

Consider estimating the population mean of a response variable when the responses could be missing depending on a covariate. This type of missing data commonly arises in survey questionnaires conducted in many areas of applied science, and nonresponses to ambiguous cases could lead to biased inference without effective correction by randomization (Cochran 1977). It also occurs when a double sampling design is used to omit some responses due to demographic constraints (Neyman 1938). If the missing mechanism is not completely random, a measure of its dependence on other covariates can be used to impute values for the nonresponses to rectify the potentially biased inference. Statistical inference with partial nonresponse data has been widely discussed since the early work by Yates (1933), Anderson (1957), Orchard and Woodury (1972), see for example, Little and Rubin (2002). Analysis with data missing at random (MAR, Rubin 1976), and the EM algorithm (Demp-

Published online: 29 December 2010

ster et al. 1977) have been extensively used. The MAR condition is a basic assumption upon which most parametric and semiparametric inference with missing data have been developed.

Suppose a random sample with incomplete responses and complete covariates is observed from a double sampling design, and denoted by

$$(X_i, Y_i, \delta_i), \quad i = 1, 2, \ldots, n. \tag{1.1}$$

All the covariates X_i are observed, and $\delta_i = 1$ if Y_i is observed, otherwise $\delta_i = 0$. Suppose that the mean of Y, $\mu = EY$, would be estimated under the assumption of MAR, that is, missing Y depends mainly on the covariate X

$$P(\delta = 1|X, Y) = P(\delta = 1|X) = p(X). \tag{1.2}$$

The missing pattern function $p(x)$ defined under MAR is an analog of the well-known propensity score $p(x, \alpha)$ of parametric inference, which is traditionally termed as "propensity to be exposed to a treatment" (Rosenbaum and Rubin 1983). Without assuming a parametric likelihood model or a parametric regression model together with a propensity model, a nonparametric approach to estimating μ depends on effective estimation of both the regression function $m(x) = E(Y|x)$ and the propensity score $p(x)$, such that proper imputation can be used to make up for the loss of incomplete data information.

There are two basic approaches to nonparametric imputation. The nonparametric regression weighted estimation and the classical Horvitz-Thompson (HT) inverse weighting estimation. The regression method imputes a missing value $(Y_i, \delta_i = 0)$ with a weighted regression estimate for $m(X_i)$, and also a similarly weighted estimate for $p(X_i)$. The HT method weights each observed response $(Y_i, \delta_i = 1)$ by the inverse of an estimated probability of observation to reflect a proper sample size. Under MAR, the estimated probabilities are the estimated propensity scores $p(X_i)$ based on the nonparametric inference frame (1.1) and (1.2).

A basic nonparametric regression imputation is the kernel-weighted regression (KR) estimator introduced by Cheng and Wei (1986):

$$\begin{aligned}\tilde{\mu} &= \frac{1}{n}\sum_{i=1}^{n}\tilde{m}(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{n} W_h(X_i, X_j)\delta_j Y_j \Big/ \sum_{j=1}^{n} W_h(X_i, X_j)\delta_j\right\}.\end{aligned} \tag{1.3}$$

The regression function is estimated by $\tilde{m}(x)$, W is a symmetric probability density function (pdf), and $W_h(u, x) = h^{-1}W((u - x)/h)$. An analog of (1.3) is

$$\mu_{KR} = \frac{1}{n}\sum_{i=1}^{n}\{\delta_i Y_i + (1 - \delta_i)\tilde{m}(X_i)\}. \tag{1.4}$$

Estimators (1.3) and (1.4) were proved to be asymptotically equivalent as they approximate the same normal distribution under the same regularity conditions on the regression function, the propensity score and the kernel bandwidth h (Cheng 1994). This asymptotic normality has also been proved using the empirical likelihood approach, see for example, Wang and Rao (2002).

The idea of using kernel regression weights was also applied to define nearest neighbor (NN) regression weights, for example, Cheng (1984, 1994). For a finite positive integer K, an NN imputation estimator is defined as

$$\mu_{NN} = \frac{1}{n}\sum_{i=1}^{n}\{\delta_i Y_i + (1 - \delta_i)m_K(X_i)\}. \tag{1.5}$$

Here $m_K(X_i) = \frac{1}{K}\sum_{j=1}^{K} Y_{i(j)}$, and $\{(X_{i(j)}, Y_{i(j)}) : \delta_{i(j)} = 1, j = 1, \ldots, K\}$ is a set of K observed data pairs, and $\boldsymbol{X}_{i(j)}$ denotes the jth nearest neighbor to X_i among all the covariates X's corresponding to those Y_k's with $\delta_k = 1$. The imputed kernel estimates $\tilde{m}(X_i)$ of (1.4) are replaced by the nearest-neighbors estimate $m_K(X_i)$ in (1.5), and the kernel bandwidth h is replaced by a random distance defined between the covariates.

The classical HT weighting scheme recovers the incomplete data information by inverting the sampling weights to reflect the effective sample size. Under MAR, a basic HT imputation estimator of μ is defined by inverting the estimated propensity score:

$$\mu_{HT} = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i Y_i}{w_i}, \tag{1.6}$$

where $w_i = \tilde{p}(X_i) = \sum_{j=1}^{n}\delta_j W_h(X_j, X_i)/\sum_{j=1}^{n} W_h(X_j, X_i)$ estimates the propensity score $p(X_i)$ using the same kernel smoothed estimate as defined by (1.3). Alternatively, the sample size n in (1.6) can be replaced by an adjusted total, that is, a ratio estimate of the effective sample size. This yields the commonly-used HT ratio estimator:

$$\mu_{HTR} = \left(\sum_{i=1}^{n}\delta_i Y_i/w_i\right) \Big/ \left(\sum_{i=1}^{n}\delta_i/w_i\right). \tag{1.7}$$

The ratio estimator is generally preferred to the naive estimator with complete data and, whether any difference could exist with the analysis of missing data will be examined later in a simulation study.

It is notable that Robins et al. (1994) used the HT inverse probability weighting to estimate a semiparametric regression function $m(x, \beta)$ when some covariates are missing at random. A remarkable advantage is that the method is

asymptotically efficient, when either the parametric regression model $m(x, \beta)$ or the propensity score model $p(x, \alpha)$ is correctly specified. This is termed the double-robustness (DR) property by Scharfstein et al. (1999), and it has been extensively used with semiparametric inference. Thereafter, the DR property has attracted much discussion, for example, Robins and Rotnitzky (2001), Carpenter et al. (2006), Kang and Schafer (2007), and Qin et al. (2008). Kang and Schafer (2007) questioned whether the DR property could be lost when both models are not correctly specified. Nevertheless, the possible failure of consistent estimation was remarked when the inverse probability weights are highly variable (Robins et al. 2007).

In contrast, a nonparametric analog of the DR property, coined the Robins-Rotnitzky-Zhao estimator, was formulated by Carpenter et al. (2006, formula (5)), and also by Qin et al. (2008, p. 798). Without parametric modeling, a nonparametric analog of the DR property requires both the regression function and the propensity score be ideally smooth functions. A polynomial regression function and a logistic-type propensity score have been used in many simulation studies under semiparametric modeling. This modifies the basic HT estimator (1.6) by defining a nonparametric doubly-robust HT estimator as

$$\mu_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left[\tilde{m}(X_i) + \frac{\delta_i\{Y_i - \tilde{m}(X_i)\}}{w_i}\right], \tag{1.8}$$

where $\tilde{m}$ and w_i are the estimates of $m(x)$ and $p(x)$ in (1.3) and (1.6), respectively. Clearly, estimator μ_{DR} modifies $\tilde{\mu}$ of (1.3) by using inversely weighted regression residuals. This motivates the definition of a nonparametric DR property in the sense that the nonparametric DR estimator μ_{DR} can be efficient so long as either the regression function $m(x)$ or the propensity score $p(x)$ is sufficiently smooth. When both $m(x)$ and $p(x)$ are ideally smooth, it can be expected, as evidenced in a simulation study, that all the imputation estimators using the same kernel-estimated weights yield comparable performance to μ_{DR} in terms of the bias, the MSE, a standardized z-score and the coverage probability of confidence intervals. In this case, the NN imputation estimator μ_{NN} could yield larger sampling variance, hence larger MSE when using a smaller K for convenience. In view of the previous remark on the failure of the parametric DR property (Robins et al. 2007), it is of interest whether the nonparametric estimator μ_{DR} is surely preferred to the NN estimator μ_{NN} when both $m(x)$ and $p(x)$ are smooth and well estimated by the kernel method. This will be examined using a comprehensive simulation study in Sect. 3.

This study will address two issues. The first goal is to prove a basic asymptotic normality for the NN imputation using a multivariate covariate, while similar results in the literature were given with a univariate covariate, e.g., Shao and Wang (2008). Section 2 proves Theorem 1 that distinct normal approximations for the NN and the KR imputation are achieved under identical regularity conditions. This characterizes the difference in the asymptotic variance between the two local-weighting schemes that has not been well illustrated in the literature. The case 1 simulation study of Sect. 3 calibrates this difference, and shows that all the methods except the NN are comparably efficient under ideally regular conditions. In contrast to Theorem 1, the second goal of this study is to examine the difference in performance between the NN imputation and the other imputation methods when the regularity conditions of Theorem 1 are not satisfied. From the simulation study case 2, it is found that among all imputation methods using kernel estimated weights, the DR estimator μ_{DR} and a nearest-neighbors (NN) modified estimator μ_{DR2} (to be defined in Sect. 3) yield the most stable performance in terms of the MSE and the z-score, being insensitive to discontinuity in the propensity score. When the propensity scores vary widely across the mixture distribution of the covariate as in the case 3 simulation, estimators using the NN imputation with a small K, 1 or 2, become highly competitive. They yield the best performance in terms of smaller MSE and more accurate coverage probability of confidence interval for the true mean, while the NN-modified estimator μ_{DR2} becomes the second best in the same simulation study. Section 4 will apply the imputation estimators to two datasets using simulated MAR designs. For the orthodontic growth dataset of size 27 (Potthoff and Roy 1964), a miniature simulation study was conducted to resemble the case 3 simulation study, yielding expected computational results. For the iris flower dataset (cf. Fisher 1936), similar simulations were conducted for studying the accuracy in classifying the three species and for estimating the species proportions. The NN imputation method obtains better classification accuracy than the KR method does under general missing data mechanisms. In both empirical studies, the NN imputation with a small K, 1 or 2, presented the best performance. It is well-known that the NN method differs from the others in using a random distance as a variable bandwidth defined by the covariate distribution instead of a constant bandwidth. While a proof for a general joint distribution under the MAR condition is beyond the scope of this study, it is examined by computation that the NN imputation could be less sensitive to the variation in the propensity scores and the unknown covariate distribution.

2 Nearest neighbor imputation

The K-nearest neighbor (K-NN) decision rule due to Fix and Hodges (1951) has been widely used in pattern recognition. Logtsgaarden and Quesenberry (1965) applied it to yield consistent estimation of a probability density function,

and Cover and Hart (1967) discussed admissibility of 1-NN classification rule. In various scientific computing environments, K-NN estimation has been widely used to study classification with multivariate data. Nearest neighbor rules in statistical estimation were discussed with hot-deck imputation (Sande 1979), and nonparametric regression (Cheng 1984). In a study of the KR imputation, the use of K-NN imputation was remarked by Cheng (1994). Methods of NN imputation were also studied by Lee et al. (1994), Rancourt (1999), Chen and Shao (2000, 2001), and Shao and Wang (2008). Most of these studies discussed missing responses in a nonparametric or semiparametric regression model with a 1-dimensional covariate X.

The KR imputation estimator (1.4) and the NN estimator (1.5) are constructed by locally-weighted nonparametric regression, but differ in the statistical distance used. With small or moderate sample size n, the KR imputation may find difficulty in using the local-bandwidth weighting with sparse high-dimensional data. In contrast, the NN imputation uses a random statistical distance between the covariates. Thus, the NN rule is basically unaffected by discontinuity of $p(x)$, sparse data or multi-dimensional covariate X. Computations for such cases will be illustrated in the simulation study of Sect. 3.

Asymptotic normality for the NN method has not been fully discussed when the data are incomplete. It is known that like KR estimation using a small bandwidth, the estimator μ_{NN} using a small K would yield negligible bias in estimating μ, but a larger variance. In theory, asymptotic variance can only be derived under regularity conditions (ideal smoothness) of the regression function $m(x)$ and the propensity score $p(x)$. Such asymptotic properties under regularity conditions are typical facts of the kernel estimator μ_{KR}, and the HT type estimators μ_{HTR} and μ_{DR}, but not well studied for the NN estimator μ_{NN} in the literature. Compared to the asymptotic normality of the KR estimator μ_{KR} (Cheng 1994), a less asymptotically efficient result for the NN estimator μ_{NN} is obtained under essentially the same regularity conditions of the joint distribution. This is given in Theorem 1 below.

Theorem 1 *Assume* $EY^2 < \infty$ *and that the conditional variance function* $\sigma^2(x) = \mathrm{Var}(Y|x)$, *the regression function* $m(x)$, *and the propensity score* $p(x)$ *are finite and first-order differentiable. Then, the NN imputation estimator* μ_{NN} *of* (1.5) *yields the approximation in distribution*:

$$\sqrt{n}(\mu_{NN} - \mu) \to N(0, \sigma^2(\mu_{NN})), \tag{2.1}$$

as $n \to \infty$ *where*

$$\sigma^2(\mu_{NN}) = \mathrm{Var}(Y) + \left(1 + \frac{1}{K}\right) E[\sigma^2(X)(1 - p(X))] + E\left[\frac{\sigma^2(X)(1 - p(X))^2}{p(X)}\right]. \tag{2.2}$$

The proof of Theorem 1 will be given in the Appendix. Under the same conditions of Theorem 1, the kernel-weighted regression imputation estimator μ_{KR} yields the asymptotic distribution $N(0, \sigma^2(\mu_{KR}))$ (Cheng 1994, Theorem 2.1), where the asymptotic variance is

$$\sigma^2(\mu_{KR}) = \mathrm{Var}(Y) + E\left[\frac{\sigma^2(X)(1 - p(X))}{p(X)}\right]. \tag{2.3}$$

The difference between the two asymptotic variances of (2.2) and (2.3) is

$$\sigma^2(\mu_{NN}) - \sigma^2(\mu_{KR}) = \frac{1}{K} E[\sigma^2(X)(1 - p(X))]. \tag{2.4}$$

Computational effects of Theorem 1 and (2.4) will be examined and illustrated in the case 1 simulation of Sect. 3. The statistical distance plays a key role in the NN imputation, allowing flexible choices such as the Euclidean distance, the Mahalanobis distance, whichever is appropriate to the data joint distribution. As a random distance between the covariate variables, it alleviates the constraint of using an optimal constant bandwidth in the KR method, particularly with a mixture, multi-dimensional or sparse distribution of the covariate X. With a small K, a K-NN imputation estimator generally yields smaller bias but larger variance compared to those of the KR imputation. Nevertheless, it usually yields smaller MSE, when the distribution of the covariate is a mixture. This will be exemplified in the case 3 simulation of Sect. 3.

It may be expected that a modification of the weights of the NN estimator μ_{NN} of (1.5), assigning unequal weights to the K nearest neighbors, for example, using a bell-shaped unimodal kernel, could possibly reduce the sampling MSE. Because no significant reduction of MSE was found from the computation in each case study, the weighted-distance NN estimation will not be discussed. In Sect. 3, three cases of simulation study are defined using smooth regression functions, and smooth or discontinuous propensity scores, and mixture covariate distributions are examined. It is remarkable that the NN imputation (using a proper K) can yield stable performance in terms of the MSE and the coverage probability of confidence interval, compared to the imputation methods using kernel-estimated weights with proper bandwidths.

3 Simulation study

Three cases of simulation study were conducted to evaluate the performance of the aforementioned imputation methods. A common regression model was used in each case

$$Y = m(X) + \varepsilon, \tag{3.1}$$

where the error variable ε is independent of X and distributed as standard normal. Varied distributions of the covariate X and the propensity score $p(X)$ were defined with the regression model (3.1). Random samples of sizes $n = 50, 100, 500$ were generated in each case. Six imputation estimators for the population mean $\mu = EY$ were computed using 1,000 replications in each case. They were compared in terms of the average bias, the MSE, the z-score (ZS denotes the ratio of the average bias to the standard error estimate), and the coverage probability of confidence intervals (denoted by CCI) for $\mu = EY$. These statistics were computed from the simulated samples and are reported in Tables 1 to 3.

1. The $K-$NN estimator μ_{NN} of (1.5) is used with $K = 1, 2, 4, 8$ in each simulation, plus two larger multiples of 4 in cases 1 and 2.
2. The kernel regression (KR) estimator is defined by (1.4) using the well-known Epanechnikov quadratic kernel function

$$W(t) = \begin{cases} 0.75(1-t^2), & \text{for } |t| \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

   When $\sum_{j=1}^{n} W_h(X_i, X_j)\delta_j = 0$ and $\tilde{p}(X_i) = 0$, there is no candidate donor within one bandwidth of the covariate X_i associated with a missing response. Then, no value would be imputed for the missing response, and the actual sample size used in this computation could be less than n.
3. The same kernel function with the selected bandwidths was used with the KR estimator μ_{KR}, the HT estimator μ_{HT}, the HT ratio estimator μ_{HTR}, and the nonparametric DR estimator μ_{DR}.
4. For the DR estimation, two imputation schemes were designed in situations where $\sum_{j=1}^{n} W_h(X_i, X_j)\delta_j = 0$: (1) as for μ_{KR}, no imputed value was used, denoted by μ_{DR}; (2) impute a value, which is the average of two observed $Y_j's$ of the two nearest covariate $X_j's$ to X_i, denoted by μ_{DR2}.

The first set of simulations, case 1 below, is used to evaluate Theorem 1, where both regression function $m(x)$ and propensity score $p(x)$ are first-order differentiable.

$$\text{Case 1} \begin{cases} m_1(x) = 2x, \\ p_1(x) = \frac{e^{2.5x}}{1+e^{2.5x}}, \\ X \sim U(0,1), \\ E(Y) = 1, \qquad E(Y_{obs}) = 1.0956, \\ P(\delta = 1) = 0.7543. \end{cases}$$

Table 1 reports the average bias, MSE, ZS and CCI out of 1,000 replications in computing the estimators for the simulation study of case 1. With smooth $m(x)$ and $p(x)$, most estimators exhibit uniformly small bias and MSE, stable CCI, giving consistent estimation with imputation, which seems to be unaffected by the slight difference between the true mean $E(Y)$ and the observed mean $E(Y_{obs})$. For a wide range of bandwidth values ($h = 0.20$, 0.15 and 0.05 corresponding to $n = 50$, 100 and 500, respectively) all bandwidth estimators yield comparable MSE and CCI, except that larger z-scores (ZS) may occur with large bandwidths or K. The NN estimator yields slightly larger MSE in general, but gives the smallest bias with smaller K such as 1 and 2. Table values for larger K such as 16 and 32 were computed to yield smaller variance or MSE, particularly for the case when the sample size is 500 or larger. Here, the asymptotic variance of (2.2) is $\sigma^2(\mu_{NN}) = 1.7005 + 0.2457/K$ which approximates $\sigma^2(\mu_{KR}) = 1.7005$, that of (2.3). It is notable in Table 1 that the MSEs of the NN method also approximate those of the KR method as K tends to infinity, in accordance with (2.4) of Theorem 1.

Next, in case 2, a quadratic polynomial $m(x)$ and a piecewise constant propensity score $p(x)$ are used. The purpose is to examine any adverse effect due to discontinuity in the propensity.

$$\text{Case 2} \begin{cases} m_2(x) = 3 - 6(x-0.6)^2, \\ p_2(x) = \begin{cases} 0.8, & 0 \le x \le 0.3, \\ 0.2, & 0.3 < x \le 0.7, \\ 0.8, & 0.7 < x \le 1, \end{cases} \\ X \sim U(0,1), \\ E(Y) = 2.44, \qquad E(Y_{obs}) = 2.26, \\ P(\delta = 1) = 0.56. \end{cases}$$

Values in Table 2 given by the simulation case 2 generally present consistent estimation. The basic HT imputation estimator μ_{HT} yields the poorest performance among all, showing that it is most sensitive to discontinuity in the propensity score $p(x)$. With an adjustment of sample size, the ratio HT imputation μ_{HTR} corrects the drawback and yields similar performance to the other estimators. The NN imputation gives satisfactory performance when K is greater than 1, and it yields the smallest bias but larger variance hence larger MSE when $K = 1$. Here, the values of ZS and CCI vary more widely than those in the ideal case 1, and are unsatisfactory with bandwidths larger than 0.20 or larger K; this is essentially due to discontinuity in the propensity score.

Case 3 differs from cases 1 and 2 by defining the covariate X as a mixture of two uniform distributions. It is designed to test the performance stability of the nonparametric imputation methods by putting heavy missingness on the data region of one component of the mixture distribution

Table 1 Average bias, n·MSE, z-score and coverage of CI for Case 1

Estimator		Sample Size n											
		50				100				500			
		Bias	n·MSE	ZS	CCI	Bias	n·MSE	ZS	CCI	Bias	n·MSE	ZS	CCI
μ_{NN}	1	−0.004	2.07	−0.63	0.954	0.003	2.03	0.74	0.956	−0.004	1.96	−1.89	0.954
	2	−0.005	2.00	−0.79	0.945	0.003	1.92	0.63	0.957	−0.003	1.73	−1.53	0.952
	4	−0.001	1.92	−0.14	0.944	0.004	1.84	0.83	0.947	−0.002	1.63	−1.33	0.955
	8	0.007	1.86	1.07	0.943	0.006	1.85	1.37	0.945	−0.003	1.60	−1.43	0.949
	16	0.027	1.90	4.41	0.946	0.014	1.85	3.23	0.954	−0.002	1.59	−1.16	0.951
	32	0.077	2.13	12.65	0.939	0.034	1.98	7.95	0.948	0.000	1.59	−0.25	0.949
μ_{KR}	0.05	0.001	1.97	0.16	0.949	0.004	1.85	0.92	0.947	−0.002	1.58	−1.14	0.953
	0.15	−0.002	1.87	−0.29	0.947	0.006	1.80	1.42	0.949	0.001	1.58	0.79	0.952
	0.20	0.000	1.84	0.06	0.945	0.009	1.80	2.13	0.95	0.004	1.58	2.36	0.95
	0.30	0.007	1.82	1.19	0.944	0.016	1.83	3.88	0.954	0.012	1.64	6.51	0.947
μ_{HT}	0.05	−0.031	1.85	−5.11	0.946	−0.006	1.79	−1.45	0.949	−0.004	1.58	−2.06	0.956
	0.15	−0.009	1.82	−1.50	0.948	0.002	1.77	0.40	0.946	−0.001	1.57	−0.41	0.955
	0.20	−0.006	1.81	−1.07	0.948	0.004	1.77	0.95	0.949	0.001	1.56	0.58	0.953
	0.30	−0.002	1.78	−0.33	0.945	0.009	1.77	2.22	0.948	0.006	1.56	3.25	0.95
μ_{HTR}	0.05	0.013	1.98	2.02	0.944	0.010	1.87	2.32	0.943	−0.001	1.59	−0.61	0.949
	0.15	0.002	1.88	0.36	0.945	0.008	1.82	1.93	0.946	0.002	1.58	1.02	0.95
	0.20	0.003	1.86	0.51	0.946	0.011	1.82	2.51	0.947	0.005	1.59	2.57	0.952
	0.30	0.009	1.83	1.41	0.945	0.017	1.84	4.05	0.951	0.012	1.64	6.63	0.945
μ_{DR}	0.05	0.000	2.00	0.06	0.951	0.004	1.87	0.90	0.95	−0.002	1.59	−1.38	0.955
	0.15	−0.004	1.90	−0.64	0.948	0.003	1.81	0.68	0.945	−0.002	1.58	−1.22	0.952
	0.20	−0.004	1.87	−0.66	0.947	0.004	1.80	0.83	0.948	−0.002	1.58	−1.00	0.953
	0.30	−0.003	1.84	−0.53	0.946	0.005	1.80	1.28	0.948	0.000	1.58	0.02	0.955
μ_{DR2}	0.05	−0.005	1.99	−0.85	0.949	0.003	1.86	0.69	0.951	−0.002	1.59	−1.38	0.955
	0.15	−0.004	1.90	−0.66	0.948	0.003	1.81	0.68	0.945	−0.002	1.58	−1.22	0.952
	0.20	−0.004	1.87	−0.65	0.947	0.004	1.80	0.83	0.948	−0.002	1.58	−1.00	0.953
	0.30	−0.003	1.84	−0.53	0.946	0.005	1.80	1.28	0.948	0.000	1.58	0.02	0.955

Table values are based on 1,000 replications. Values of K and bandwidth h are listed in the second column. ZS is the z-score of standardized average bias, CCI is the coverage probability of confidence intervals

of X. The simulation is defined as follows.

$$\text{Case 3}\begin{cases} m_3(x) = 2x + 1, \\ p_3(x) = \frac{\exp(2.5x)}{1+\exp(2.5x)}, \\ X \sim 0.3U(-3, 0) + 0.7U(0, 4), \\ E(Y) = 2.9, \qquad E(Y_{obs}) = 5.009, \\ P(\delta = 1) = 0.679. \end{cases}$$

Compared to the previous cases 1 and 2, the highly unbalanced missing pattern $p(x)$ in the mixture distribution of the covariate appears to present serious adverse effect on the estimators using kernel bandwidths. Consistent estimation may not be secured if the values n·MSE increase steadily with n. In this case, estimators μ_{HT}, μ_{KR}, μ_{HTR}, and μ_{DR} tend to give larger average bias, MSE and ZS, but smaller CCI as compared to the estimators μ_{NN} and μ_{DR2}. It is seen from Table 3 that the NN estimator μ_{NN} with $K = 1, 2$ yields the best results, comparable to the estimator μ_{DR2}. Clearly, μ_{DR2} costs more computation time in requiring a bandwidth h that roughly decreases from 1.5 to the range $(0.5, 0.8)$ as the sample size n increases from 50 to 500.

With unbalanced missingness in the mixture distribution of the covariate, it is natural to consider using two different bandwidths, that is, a larger one for the sparse data region, and a smaller one for the less-missing part. By using various combinations of two bandwidths h and two Ks in case 3, a modified simulation study indicates that there are no ap-

Table 2 Average bias, n·MSE, z-score and coverage of CI for Case 2

Estimator		Sample Size n											
		50				100				500			
		Bias	n·MSE	ZS	CCI	Bias	n·MSE	ZS	CCI	Bias	n·MSE	ZS	CCI
μ_{NN}	1	0.006	3.55	0.72	0.953	−0.002	4.38	−0.28	0.952	0.004	4.10	1.41	0.954
	2	−0.007	2.89	−0.96	0.952	−0.007	3.64	−1.09	0.949	0.003	3.46	1.16	0.953
	4	−0.021	2.54	−2.98	0.95	−0.014	3.14	−2.48	0.956	0.003	3.11	1.08	0.954
	8	−0.055	2.47	−8.13	0.939	−0.032	2.86	−6.13	0.952	0.001	2.86	0.59	0.955
	16	−0.098	2.79	−14.38	0.92	−0.065	2.92	−13.02	0.935	−0.004	2.71	−1.51	0.959
	32	−0.165	3.61	−24.70	0.89	−0.110	3.73	−21.88	0.903	−0.017	2.66	−7.63	0.945
μ_{KR}	0.10	−0.012	2.93	−1.56	0.948	−0.010	3.32	−1.78	0.949	−0.006	2.69	−2.47	0.953
	0.15	−0.012	2.73	−1.64	0.955	−0.020	3.02	−3.63	0.947	−0.016	2.64	−7.16	0.943
	0.20	−0.025	2.53	−3.48	0.953	−0.034	2.83	−6.45	0.948	−0.030	2.82	−13.72	0.926
	0.25	−0.041	2.44	−5.95	0.942	−0.051	2.80	−10.01	0.938	−0.047	3.38	−21.81	0.885
	0.30	−0.059	2.45	−8.67	0.937	−0.069	2.92	−13.94	0.927	−0.065	4.36	−30.63	0.84
μ_{HT}	0.10	−0.277	8.54	−28.58	0.855	−0.162	7.10	−24.25	0.893	−0.080	6.14	−32.97	0.818
	0.15	−0.212	6.52	−22.87	0.899	−0.154	6.12	−25.17	0.889	−0.114	9.31	−48.39	0.654
	0.20	−0.213	5.97	−24.70	0.886	−0.180	6.72	−30.37	0.846	−0.153	14.60	−63.40	0.5
	0.25	−0.224	5.77	−27.84	0.873	−0.208	7.52	−36.70	0.8	−0.189	20.86	−77.37	0.315
	0.30	−0.223	5.40	−29.14	0.863	−0.217	7.65	−40.04	0.775	−0.203	23.52	−84.95	0.246
μ_{HTR}	0.10	−0.037	2.96	−4.83	0.944	−0.027	3.42	−4.66	0.945	−0.012	2.84	−5.09	0.949
	0.15	−0.033	2.86	−4.43	0.953	−0.031	3.21	−5.47	0.947	−0.023	2.86	−9.87	0.934
	0.20	−0.042	2.71	−5.80	0.951	−0.042	3.05	−7.91	0.946	−0.036	3.15	−16.37	0.917
	0.25	−0.054	2.59	−7.79	0.942	−0.059	2.99	−11.40	0.939	−0.054	3.81	−24.61	0.872
	0.30	−0.069	2.56	−10.11	0.937	−0.076	3.06	−15.29	0.923	−0.072	4.85	−33.52	0.81
μ_{DR}	0.10	−0.008	3.06	−1.06	0.945	−0.005	3.58	−0.87	0.952	0.001	2.85	0.35	0.956
	0.15	−0.001	2.98	−0.08	0.956	−0.007	3.34	−1.24	0.95	−0.001	2.71	−0.58	0.953
	0.20	−0.003	2.79	−0.35	0.955	−0.010	3.11	−1.77	0.947	−0.004	2.63	−1.84	0.953
	0.25	−0.009	2.63	−1.19	0.96	−0.015	2.91	−2.87	0.947	−0.010	2.56	−4.34	0.95
	0.30	−0.019	2.50	−2.67	0.951	−0.026	2.77	−4.93	0.941	−0.021	2.62	−9.34	0.937
μ_{DR2}	0.10	0.001	3.12	0.17	0.954	−0.004	3.60	−0.69	0.954	0.001	2.85	0.35	0.956
	0.15	0.001	2.99	0.14	0.956	−0.007	3.34	−1.21	0.95	−0.001	2.71	−0.58	0.953
	0.20	−0.002	2.79	−0.30	0.955	−0.010	3.11	−1.77	0.947	−0.004	2.63	−1.84	0.953
	0.25	−0.009	2.63	−1.19	0.96	−0.015	2.91	−2.87	0.947	−0.010	2.56	−4.34	0.95
	0.30	−0.019	2.50	−2.67	0.951	−0.026	2.77	−4.93	0.941	−0.021	2.62	−9.34	0.937

Table values are based on 1,000 replications. Values of K and bandwidth h are listed in the second column. ZS is the z-score of standardized average bias, CCI is the coverage probability of confidence intervals

parent improvements. Details are given in the manuscript on the author's website www.stat.sinica.edu.tw/pcheng/.

4 Empirical study

Two datasets with small to moderate sample sizes will be studied in this section. The first study examines the performance of the imputation estimators with the small orthodontic growth dataset where Theorem 1 and related asymptotic properties are invalid. For estimating a mean parameter from this small dataset, the NN and the NN-modified DR estimators are found to be most effective among all the imputation estimators under artificial MAR designs. The second study uses the iris flower data (in the UCI Machine Learning Repository, MLR) and considers classifying the species and estimating the proportions. It is seen that all the imputation estimators can be used to estimate the proportions, but

Table 3 Average bias, n·MSE, z-score and coverage of CI for Case 3

	K/h	Sample Size n											
		50				100				500			
		Bias	n·MSE	ZS	CCI	Bias	n·MSE	ZS	CCI	Bias	n·MSE	ZS	CCI
μ_{NN}	1	0.684	41.78	35.70	0.811	0.553	54.06	36.09	0.797	0.282	80.82	31.15	0.842
	2	0.784	45.55	45.49	0.704	0.648	61.45	46.37	0.688	0.345	90.25	44.15	0.713
	4	0.920	54.79	58.20	0.549	0.772	75.89	60.44	0.519	0.434	117.90	62.99	0.500
	8	1.129	74.95	75.43	0.359	0.921	98.32	79.06	0.290	0.550	170.15	89.08	0.192
μ_{KR}	0.8	1.405	118.21	71.00	0.376	1.217	175.91	72.97	0.355	0.778	377.77	63.62	0.442
	1.0	1.331	108.75	66.29	0.435	1.144	159.64	67.54	0.423	0.721	327.86	61.63	0.475
	1.4	1.210	93.53	60.09	0.513	1.036	135.09	62.29	0.491	0.659	265.05	67.51	0.446
	1.8	1.120	80.98	58.58	0.539	0.970	118.17	62.38	0.499	0.683	260.17	92.74	0.160
	2.0	1.094	77.24	58.55	0.555	0.955	112.75	64.85	0.482	0.728	284.89	116.03	0.038
μ_{HT}	0.8	0.583	27.45	40.32	0.769	0.581	45.94	52.55	0.622	0.485	140.45	71.61	0.334
	1.0	0.587	27.80	40.39	0.767	0.583	46.46	52.21	0.633	0.495	146.34	71.80	0.328
	1.4	0.605	28.79	41.65	0.753	0.604	49.02	53.80	0.608	0.543	168.52	83.98	0.205
	1.8	0.634	30.33	44.35	0.739	0.640	52.80	58.74	0.540	0.605	198.95	107.15	0.087
	2.0	0.653	31.28	46.17	0.721	0.661	55.13	61.73	0.506	0.636	215.87	121.32	0.041
μ_{HTR}	0.8	1.600	145.57	85.48	0.225	1.459	238.55	90.86	0.182	1.111	701.84	85.24	0.223
	1.0	1.584	143.76	82.62	0.245	1.454	238.95	87.58	0.211	1.156	759.66	85.69	0.217
	1.4	1.585	144.19	82.07	0.244	1.493	250.25	90.05	0.178	1.300	923.21	103.54	0.121
	1.8	1.611	147.01	86.66	0.206	1.552	264.00	101.92	0.101	1.444	1094.35	142.20	0.028
	2.0	1.626	148.49	90.12	0.183	1.580	270.44	109.28	0.067	1.503	1169.48	169.22	0.008
μ_{DR}	0.8	1.381	115.33	69.17	0.397	1.184	168.67	70.10	0.385	0.720	336.82	57.80	0.513
	1.0	1.293	104.48	63.27	0.468	1.094	149.65	63.29	0.470	0.637	277.01	52.32	0.601
	1.4	1.139	86.68	54.50	0.580	0.950	120.65	54.53	0.584	0.536	201.98	49.54	0.655
	1.8	1.016	72.10	50.17	0.638	0.851	100.41	50.91	0.636	0.534	178.70	62.74	0.505
	2.0	0.975	67.29	49.02	0.641	0.823	93.08	51.65	0.631	0.570	189.59	77.60	0.304
μ_{DR2}	0.8	0.742	43.50	41.57	0.746	0.612	58.29	42.31	0.733	0.320	85.45	38.62	0.781
	1.0	0.738	43.35	41.11	0.747	0.611	58.36	42.18	0.738	0.332	89.34	40.01	0.759
	1.4	0.739	43.59	40.93	0.752	0.626	60.13	43.26	0.723	0.389	107.66	48.72	0.666
	1.8	0.757	44.38	42.61	0.734	0.666	64.26	47.08	0.683	0.490	146.19	68.05	0.415
	2.0	0.775	45.36	44.22	0.714	0.696	67.31	50.55	0.642	0.553	175.45	82.09	0.249

Table values are based on 1,000 replications. Values of K and bandwidth h are listed in the second column. ZS is the z-score of standardized average bias, CCI is the coverage probability of confidence intervals

only the KR and the NN estimators can be used to classify the iris species under various missing data mechanisms.

4.1 Orthodontic growth data

The data in Table 4 (Potthoff and Roy 1964) are orthodontic growth measurements for 11 girls and 16 boys. For each subject, the distance from the center of the pituitary to the maxillary fissure was recorded at the ages of 8, 10, 12, and 14 years. Assuming the four distance measures were observed from a multivariate normal distribution, Little and Rubin (2002, Table 11.4) examined the inference for the linear regression parameters as if the data were incomplete. An artificial deletion mechanism was designed to be MAR, specifically, values at age 10 years were deleted for cases with low values at age 8 years. By analogy with their MAR design, we may assume that the finite population mean of the 27 measures of all the boys and girls at age 14 is to be estimated using a similar MAR design. Thus, measures at age 14 are the response Y values, and let those at age 12 be defined as the covariate X values. Our goal is to examine the performance of the imputation methods in estimating $\mu = EY = 26.09$.

In accordance with formulae (1.1) and (1.2), let some Y values be deleted according to the propensity score $p(x)$ defined as

$$p(x) = \begin{cases} 0.9, & x < 25, \\ 0.4, & x \geq 25. \end{cases} \tag{4.1}$$

Because of the small data size, $n = 27$, this deletion mechanism was only simulated 20 times. A typical simulated missing data is presented in Table 4, where the deleted Y values are quoted in parentheses. For these twenty simulated datasets, all of the previously discussed imputation methods are computed and the results are summarized in Table 5. As in the previous simulation study of case 3, there was no clear advantage in using pairs of bandwidths h, one for the girls and another for the boys, or pairs of K, when both observed x and y values were not sparsely distributed. Thus, the results in Table 5 were reported using a single h and K for this simulated missing data analysis. By missing a large proportion of the response values of the boys, but not of the girls, it is expected that the results in Table 5 would resemble those in Table 3. Indeed, the NN estimator μ_{NN}, with $K = 1, 2$, and the NN-modified estimator μ_{DR2} yield the best results among all estimators.

Table 4 Orthodontic growth data for 11 girls and 16 boys

Girl	Age (in years)		Boy	Age (in years)	
	12(X)	14(Y)		12(X)	14(Y)
1	21.5	23.0	1	29.0	(31.0)
2	24.0	25.5	2	23.0	26.5
3	24.5	26.0	3	24.0	27.5
4	25.0	(26.5)	4	26.5	(27.0)
5	22.5	23.5	5	22.5	(26.0)
6	21.0	22.5	6	27.0	28.5
7	23.0	25.0	7	24.5	26.5
8	23.5	24.0	8	24.5	25.5
9	22.0	(21.5)	9	31.0	26.0
10	19.0	19.5	10	31.0	31.5
11	28.0	(28.0)	11	23.5	(25.0)
			12	24.0	28.0
			13	26.0	29.5
			14	25.5	26.0
			15	26.0	(30.0)
			16	23.5	25.0

Sources: Potthoff and Roy (1964); Little and Rubin (2002)

4.2 Iris flower data

The iris flower dataset is well known from the study of classification using linear discriminant analysis (Fisher 1936). The dataset consists of 50 samples from each of three species of iris flowers (iris setosa, iris virginica and iris versicolor). Four features were measured from each sample, they are the length and the width of sepal and petal, in centimeters. Scatter plots of the six paired features are available in the MLR. The plots indicate that iris setosa is linearly separable from the other two species, which are not separable from each other. The attribute of interest is the species indicator variable, denoted by Y, say, $Y = 0$ stands for iris setosa, $Y = 1$ for virginica and $Y = 2$ for versicolor. The useful covariate is the predictor vector $\boldsymbol{X} = (X_1, X_2, X_3, X_4)$ of the four features (Sepal length, Sepal width, Petal length, Petal width). In the context of machine learning, data measurements of validation samples are used to supervise the training of a classifier, and then the unknown attributes (species, response types) of the remaining test samples are classified as if they were missing completely at random (MCAR). It is however notable that both the attributes and the covariates in specific data regions are used in training the classifiers, this is called feature extraction in supervised learning. As a consequence of repeated sampling and training on the same data, such missing data mechanisms may not be MCAR or MAR, but rather missing not at random (MNAR), which is also termed non-ignorable missing in

Table 5 A simulated incomplete growth data analysis for estimating EY

	K/h	Bias	Var	n·MSE		h	Bias	Var	n·MSE		h	Bias	Var	n·MSE
μ_{NN}	1	−0.133	0.093	2.87	μ_{HT}	2.0	−2.757	4.300	315.57	μ_{DR}	2.0	−0.256	0.084	3.92
	2	−0.182	0.071	2.71		2.1	−2.666	3.988	294.19		2.1	−0.192	0.104	3.66
	4	−0.282	0.079	4.19		2.2	−2.559	3.826	274.94		2.2	−0.191	0.102	3.60
	8	−0.367	0.080	5.68		2.5	−2.264	3.658	232.20		2.5	−0.189	0.099	3.51
μ_{KR}	2.0	−0.284	0.079	4.20	μ_{HTR}	2.0	−0.321	0.109	5.60	μ_{DR2}	2.0	−0.160	0.085	2.88
	2.1	−0.225	0.098	3.88		2.1	−0.322	0.099	5.35		2.1	−0.141	0.092	2.90
	2.2	−0.227	0.097	3.89		2.2	−0.321	0.093	5.16		2.2	−0.140	0.090	2.83
	2.5	−0.233	0.097	3.95		2.5	−0.313	0.083	4.78		2.5	−0.138	0.087	2.74

Table values are based on 20 simulated datasets. Numbers of nearest neighbors are denoted by K, and h is the kernel bandwidth

Table 6 A classification analysis for the iris flower data

	Propensity score	Average misclassification No.		Average missing No.
		KR/$h = 0.9$	1-NN	
MAR	(0.7, 0.1)	5.0	3.0	100.8
	(0.1, 0.7)	7.8	4.4	79.2
	(0.7, 0.3)	3.7	2.7	82.2
	(0.3, 0.7)	4.7	3.9	67.6
	(0.9, 0.4)	2.2	1.5	61.8
	(0.4, 0.9)	3.1	3.0	43.8
	(0.6, 0.4)	3.8	3.2	79.0
	(0.4, 0.6)	4.4	3.9	71.9
	(0.5, 0.5)	4.0	3.6	75.3
MCAR	0.4	5.5	4.4	90.1
	0.5	4.0	3.5	75.2
	0.6	3.1	2.7	59.9

Propensity scores are given by formula (4.2) under MAR, and are constants under MCAR. Table values are based on 500 replications

parametric inference (Little and Rubin 2002). Without supervised learning, the covariate $\boldsymbol{X}$ is used for selecting a validation sample at random, then the attributes of the test sample are regarded as MAR, or MCAR if the sampling is independent of $\boldsymbol{X}$. In this study, classification accuracy for the iris flower species is examined using the KR and the NN imputation methods, and compared under MAR and MCAR designs. Meanwhile, it is notable that the HT-type estimators can by definition only estimate the species proportions or the species total counts, the latter take the same value of 50 for each of the iris species.

The test accuracy of the KR and the NN imputation methods can be assessed and compared under both MAR and MCAR designs. The MCAR design can be simply defined using a constant propensity score $p(\boldsymbol{x})$ as shown in Table 6. A typical MAR design on the attribute Y is defined and simulated with a propensity score, for example,

$$p(\boldsymbol{x}) = \begin{cases} 0.7, & x_2 < 3.0, \\ 0.1, & x_2 \geq 3.0. \end{cases} \tag{4.2}$$

The feature x_2, the Sepal width, was chosen as the covariate because its values spread across the three species more evenly than the other three features. The simulation was repeated 500 times in accordance with the data size 150, and the average observed sample size is 49. For each simulated sample, missing attributes were imputed using the KR estimator μ_{KR} and the NN estimator μ_{NN}, respectively, and checked against the true attributes. The uniform density function was used as the kernel for ease of computation. By using the feature x_2, the bandwidth $h = 0.9$ and $K = 1$ yield the least average number of test errors for the estimators μ_{KR} and μ_{NN}, respectively. The average misclassified counts are reported in Table 6, where results obtained from other propensity scores are also listed under both MAR and MCAR designs. The table values indicate that the NN method yields better classification accuracy of the iris flower species than the KR method does. Under the MAR and MCAR designs, the ranges of average misclassified errors can be wider than those obtained from the analysis using a support vector machine with supervised learning, see for example, Gunn (1998).

Meanwhile, in the same simulation study, all the imputation estimators were used to estimate the same total count of 50 for each of the iris species, ignoring the data size of 150. The average bias, variance and MSE out of 500 replications are reported in Table 7. The NN estimator μ_{NN} using $K = 1$, the 1-NN estimator, yields the best performance among all. Imputation estimators μ_{KR}, μ_{DR}, and μ_{DR1} yield comparable results using bandwidths in the range [1.0, 1.5], which are better with smaller bandwidths close to 1.0. The HT estimators μ_{HT} and μ_{HTR} yield poor results using the bandwidth 1.5, in particular, the biases of the estimator μ_{HT} may not sum to zero, but μ_{HTR} rectifies the drawback using a proper range of bandwidth. It is worth noting that the 1-NN and the NN-modified DR estimator μ_{DR1} (using the first nearest neighbor) can estimate the total count of iris setosa without error. This happens even though the scatter plots (in the MLR) show that iris setosa is linearly separable from the other species either by X_3, the Petal length, or by X_4, the Petal width; but not by X_2, the Sepal width, which is used to define the MAR design of this comparison study.

In summary, it is shown that under the MAR design the 1-NN imputation estimator presents the best performance of both species classification and proportion estimation for the iris data.

5 Discussion

Nearest neighbor estimation has been widely used in studying classification and discrimination with multivariate data where the source of information is often presented as a mixture distribution. This study examines the performance of the NN imputation method in estimating a population mean of incomplete responses and also in classifying the incomplete responses which are missing at random depending on the covariate.

For estimating a population mean, Theorem 1 and simulation case 1 clarify the computational difference between the NN imputation and the KR imputation in terms of distinct asymptotic normality properties which hold under ideal regularity conditions on the regression function and the

Table 7 Average bias, variance and MSE in estimating the same total count of the iris species

	K/h	Bias			Variance			MSE		
		setosa	virginica	versicolor	setosa	virginica	versicolor	setosa	virginica	versicolor
μ_{NN}	1	0.000	−0.306	0.306	0.000	3.515	3.515	0.000	3.602	3.602
	2	−0.001	−0.143	0.144	0.001	4.400	4.400	0.001	4.412	4.412
	4	−1.285	0.934	0.351	20.547	25.108	4.977	22.156	25.929	5.091
	8	−11.322	10.557	0.764	96.856	109.054	8.279	224.844	220.292	8.847
μ_{KR}	0.5	−1.679	8.253	−6.574	43.329	25.391	22.177	46.061	93.452	65.349
	1.0	0.072	−0.003	−0.069	1.822	8.260	8.501	1.824	8.244	8.488
	1.5	−0.032	1.452	−1.420	0.050	13.550	13.448	0.051	15.631	15.438
	1.8	−0.477	3.727	−3.250	0.161	16.297	15.971	0.388	30.158	26.503
	2.0	−2.194	6.321	−4.127	2.372	19.897	16.940	7.180	59.812	33.941
μ_{HT}	0.5	−7.837	−5.032	−14.886	124.265	15.764	28.739	185.431	41.049	250.290
	1.0	1.128	−2.704	−3.666	52.068	6.463	21.243	53.235	13.764	34.643
	1.5	1.175	−0.330	−1.651	30.708	9.193	14.865	32.027	9.283	17.560
	1.8	0.234	1.794	−2.708	16.748	11.079	18.925	16.769	14.275	26.219
	2.0	−2.051	3.298	−3.874	8.061	12.270	18.709	12.253	23.120	33.678
μ_{HTR}	0.5	0.999	5.711	−6.711	93.006	47.977	39.562	93.818	80.500	84.515
	1.0	2.782	−0.831	−1.951	26.149	14.113	17.489	33.838	14.776	21.260
	1.5	1.341	0.012	−1.353	12.094	12.890	15.051	13.868	12.864	16.852
	1.8	0.415	2.080	−2.495	6.648	13.568	16.691	6.807	17.866	22.882
	2.0	−1.216	4.284	−3.068	5.076	14.844	16.608	6.545	33.169	25.989
μ_{DR}	0.5	−1.703	8.173	−6.471	43.884	25.453	22.494	46.695	92.207	64.322
	1.0	0.072	−1.167	1.095	1.822	6.484	6.714	1.824	7.835	7.901
	1.5	−0.032	−1.762	1.793	0.050	10.919	10.845	0.051	14.000	14.040
	1.8	−0.162	0.274	−0.112	0.437	13.939	13.173	0.462	13.986	13.159
	2.0	−0.427	1.827	−1.400	1.851	16.619	14.307	2.030	19.924	16.238
μ_{DR1}	0.5	0.000	0.631	−0.631	0.000	4.085	4.085	0.000	4.474	4.474
	1.0	0.000	−1.874	1.874	0.000	5.469	5.469	0.000	8.970	8.970
	1.5	0.000	−1.780	1.780	0.000	10.860	10.860	0.000	14.008	14.008
	1.8	−0.160	0.272	−0.113	0.431	13.922	13.184	0.456	13.969	13.170
	2.0	−0.427	1.827	−1.400	1.851	16.619	14.307	2.030	19.924	16.238

Table values are based on 500 replications. Numbers of nearest neighbors are denoted by K, and h is the kernel bandwidth

propensity score. All imputation estimators except the basic HT estimator are insensitive to discontinuity in the missing pattern as shown by the case 2 simulation study. If the propensity score is unbalanced with respect to the covariate distribution as defined in the case 3 simulation study, then only the NN estimator and the NN-modified DR estimator can yield satisfactory performance. The same advantage of these two imputation estimators is also evidenced in the empirical study of the small orthodontic growth data.

It is worth noting that classification methods using supervised learning are often evaluated when the unknown attributes of the test samples may not be MCAR or MAR, but MNAR. In this study, the KR and the NN imputation methods are applied to classify the iris flower species when the test samples are defined under both MAR and MCAR designs. This confirms once again that the NN imputation method yields the best classification accuracy of the iris flower species. In conclusion, a future study of both KR and NN imputation methods when the data are MNAR will be worthwhile.

Acknowledgements We are grateful to the editor, the associate editor and the referees for their helpful comments which have greatly improved the presentation. This study is partially supported by the NSFC (grant 10571070, to Ning) and the NSC (grant 2118-M001-010, to Cheng).

Appendix: Proof of Theorem 1

The parameter to be estimated is $\mu = E(Y)$ and the K-NN estimator μ_{NN} is defined by (1.5). The difference between the regression-type estimator μ_{NN} and μ can be expressed as

$$\begin{aligned}\mu_{NN} - \mu &= \sum_{i=1}^{n} \{\delta_i Y_i + (1-\delta_i) m_K(X_i)\}/n - \mu \\ &= R + S + T, \end{aligned} \tag{A.1}$$

where $R = \sum_{i=1}^{n} \{m(X_i) - \mu\}/n$, $S = \sum_{i=1}^{n} \delta_i \{Y_i - m(X_i)\}/n$. By first-order differentiability of the regression function $m(x)$ and the propensity function $p(x)$, a similar analysis to that for kernel regression estimation (Cheng 1994) yield that

$$T = \sum_{i=1}^{n} (1-\delta_i)\{m_K(X_i) - m(X_i)\}/n = T' + o(1/\sqrt{n})$$

asymptotically in probability, with

$$T' = \frac{1}{n} \sum_{i=1}^{n} (1-\delta_i) \left[\frac{1}{K} \sum_{j=1}^{k} \{\delta_{i(j)} Y_{i(j)} - m(X_{i(j)})\} \right]. \tag{A.2}$$

It is straightforward to see that $E(R) = E(S) = E(T') = 0$,

$$n \operatorname{Var}(R) = \operatorname{Var}(m(X)), \tag{A.3}$$

and

$$n \operatorname{Var}(S) = E[p(X)\sigma^2(X)]. \tag{A.4}$$

To compute the approximate formula of $\operatorname{Var}(T)$, or $\operatorname{Var}(T')$, it is more clear by writing the square as a product before taking the expectation:

$$\begin{aligned}(T')^2 = \frac{1}{n^2} \sum_{i}^{n} \sum_{j}^{n} (1-\delta_i)(1-\delta_j) \\ \times \left[\frac{1}{K} \sum_{k=1}^{K} \{\delta_{i(k)} Y_{i(k)} - m(X_{i(k)})\} \right] \\ \times \left[\frac{1}{K} \sum_{k'=1}^{K} \{\delta_{j(k')} Y_{j(k')} - m(X_{j(k')})\} \right]. \end{aligned} \tag{A.5}$$

In (A.5), consider for each i, $\delta_i = 0$, $i = 1, 2, \ldots, n$, the expectation of the conditional distribution of the product $\{\delta_{i(k)} Y_{i(k)} - m(X_{i(k)})\}\{\delta_{j(k')} Y_{j(k')} - m(X_{j(k')})\}$ having $\delta_j = 0$, given that $X_{i(k)} = X_{j(k')}$ and $\delta_{i(k)} = \delta_{j(k')} = 1$. This includes the identical terms having $j = i$ and those having $j \neq i$, which yield all the non-zero terms of conditional variance that are contained in the product (A.5). By smoothness of the conditional variance function $\sigma^2(x)$ and the propensity function $p(x)$, it can be derived from (A.5) that

$$\begin{aligned} n \operatorname{Var}(T') \simeq \frac{1}{K} E[\{1-p(X)\}\sigma^2(X)] \\ + E\left[\frac{\{1-p(X)\}^2 \sigma^2(X)}{p(X)} \right]. \end{aligned} \tag{A.6}$$

Formula (A.6) corrects a previous error in the same asymptotic variance parameter given in Cheng (1994, Remark 2.5). By a similar analysis to (A.6), it is straightforward to compute that

$$2n \operatorname{Cov}(S, T) \simeq 2E[\{1-p(X)\}\sigma^2(X)]. \tag{A.7}$$

Taking the sum (A.3) + (A.4) + (A.6) + (A.7), the desired asymptotic variance is approximately

$$\begin{aligned} n \operatorname{Var}(\mu_{NN}) \simeq \operatorname{Var}(Y) + \left(1 + \frac{1}{K}\right) E[\{1-p(X)\}\sigma^2(X)] \\ + E\left[\frac{\{1-p(X)\}^2 \sigma^2(X)}{p(X)} \right]. \end{aligned} \tag{A.8}$$

This verifies formula (2.2), hence (2.1), and concludes Theorem 1. It follows that the asymptotic variance of kernel mean imputation (2.3) can be reached by letting $K \to \infty$ in formula (A.8). That is, under the same regularity conditions of smoothness on the propensity and regression functions, the limiting asymptotic variance of μ_{NN} is equal to

$$\sigma^2(\mu_{KR}) = \operatorname{Var}(Y) + E\left[\frac{\{1-p(X)\}\sigma^2(X)}{p(X)} \right]. \tag{A.9}$$

References

Anderson, T.W.: Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. J. Am. Stat. Assoc. **52**, 200–203 (1957)

Carpenter, J.R., Kenward, M.G., Vansteelandt, S.: A comparison of multiple imputation and doubly robust estimation for analyses with missing data. J. R. Stat. Soc. A **69**, 571–584 (2006)

Chen, J., Shao, J.: Nearest neighbor imputation for survey data. J. Off. Stat. **16**, 113–132 (2000)

Chen, J., Shao, J.: Jacknife variance estimation for nearest neighbor imputation. J. Am. Stat. Assoc. **96**, 260–269 (2001)

Cheng, P.E.: Strong consistency of nearest neighbor regression function estimators. J. Multivar. Anal. **15**, 63–72 (1984)

Cheng, P.E.: Nonparametric estimation of mean functionals with data missing at random. J. Am. Stat. Assoc. **89**, 81–87 (1994)

Cheng, P.E., Wei, L.J.: Nonparametric inference under ignorable missing data process and treatment assignment. In: International Statistical Symposium, Taipei, vol. 1, pp. 97–112 (1986)

Cochran, W.G.: Sampling Techniques. Wiley, New York (1977)

Cover, T.M., Hart, P.E.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theory **13**, 21–27 (1967)

Dempster, A.P., Laird, N.M., Roubin, D.B.: Maximum likehood from incomplete data via the EM algorithm (with discussion). J. R. Stat. Soc. B **39**, 1–38 (1977)

Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugen., Part II **7**, 179–188 (1936)

Fix, E., Hodges, J.L. Jr.: Discriminatory analysis, nonparametric discrimination. USAF School of Aviation Medicine, Randolph Field, Tex., Project 21-49-404, Rept. 4, Contract AF41(128)-31 (1951)

Gunn, S.R.: Support vector machines for classification and regression. Technical Report MP-TR-98-05, Image Speech and Intelligent Systems Group, University of Southampton (1998)

Horvitz, D.G., Thompson, D.J.: A generalization of sampling without replacement from a finite population. J. Am. Stat. Assoc. **47**, 663–685 (1952)

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

Lee, H., Rancout, E., Sarndal, C.E.: Experiments with variance estimation from survey data with imputed values. J. Off. Stat. **10**, 231–243 (1994)

Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, New York (2002)

Logtsgaarden, D.O., Quesenberry, C.P.: A nonparametric estimate of a multivariate density function. Ann. Math. Stat. **36**, 1049–1051 (1965)

Neyman, J.: Contribution to the theory of sampling human populations. J. Am. Stat. Assoc. **33**, 101–116 (1938)

Orchard, T., Woodury, M.A.: A missing information principle: Theory and applications. In: Proc. 6th Berkeley Symposium on Math. Stat. and Prob., vol. 1, pp. 697–715 (1972)

Potthoff, R.F., Roy, S.N.: A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika **51**, 313–326 (1964)

Qin, J., Shao, J., Zhang, B.: Efficient and doubly robust imputation for covariate-dependent missing responses. J. Am. Stat. Assoc. **103**, 797–810 (2008)

Rancourt, E.: Estimation with nearest neighbor imputation at Statistics Canada. In: Proceedings of the Section on Survey Research Methods, pp. 131–138. Am. Statist. Assoc., Alexandria (1999)

Robins, J.M., Rotnitzky, A.: Comment on "Inference for semiparamentric models: some questions and an answer," by P.J. Bickel and J. Kwon. Stat. Sin. **11**, 920–936 (2001)

Robins, J.M., Rotnitzky, A., Zhao, L.P.: Estimation of regression coefficients when some regressors are not always observed. J. Am. Stat. Assoc. **89**, 846–886 (1994)

Robins, J.M., Sued, M., Quanhong, L.-G., Rotnitzky, A.: Comment: performance of double-robust estimators when inverse probability weights are highly variable. Stat. Sci. **22**, 544–559 (2007)

Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**, 41–45 (1983)

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

Sande, I.G.: A personal view of Hot Deck imputation procedures. Surv. Methodol. **5**, 238–258 (1979)

Scharfstein, D.O., Rotnitzky, A., Robins, J.M.: Adjusting for nonignorable drop-out using semiparametric nonresponse models. J. Am. Stat. Assoc. **94**, 1096–1120 (1999)

Shao, J., Wang, H.: Confidence intervals based on survey data with nearest neighbor imputation. Stat. Sin. **18**, 281–297 (2008)

Wang, Q., Rao, J.N.K.: Empirical likelihood-based inference under imputation for missing response data. Ann. Stat. **30**, 896–924 (2002)

Yates, F.: The analysis of replicated experiments when the field results are incomplete. Emporium J. Exp. Agric. **1**, 129–142 (1933)