

Statistica Sinica Preprint No: SS-2015-0204.R3

| | |
|--|---|
| Title | Convex Mixtures Imputation and Applications |
| Manuscript ID | SS-2015-0204.R3 |
| URL | http://www.stat.sinica.edu.tw/statistica/ |
| DOI | 10.5705/ss.202015.0204 |
| Complete List of Authors | Jianhui Ning Michelle Liou and Philip E. Cheng |
| Corresponding Author | Michelle Liou |
| E-mail | mliou@stat.sinica.edu.tw |
| Notice: Accepted version subject to English editing. | |

Convex Mixtures Imputation and Applications

Jianhui Ning¹, Michelle Liou², and Philip E. Cheng²

¹*School of Mathematics and Statistics, Central China Normal University*

²*Institute of Statistical Science, Academia Sinica*

Abstract: Nearest neighbor regression and kernel regression have been discussed with imputing missing data in survey sampling for decades. In this study, methods of regression imputation are examined for estimating the mean of an incomplete variable and for predicting unidentified objects in the data. Novel convex mixtures of these two regression imputation estimators are constructed for keeping stable performance when the underlying missing data conditions are non-regular in applications. Using a simulation study of two typical non-regular conditions, the mixture imputation is shown to yield improved estimation against the existing competitors. The performance of predicting unidentified classes by the convex mixtures imputation estimators is also examined using two data sets from the UCI Machine Learning Repository.

Key words and phrases: convex mixtures estimation, k-nearest neighbor imputation, kernel regression imputation, machine learning.

1. Introduction

Incomplete data commonly arise in various forms of item nonresponses in many studies using large-scale survey questionnaires. They may arise from the well-known double sampling scheme when rarely observed responses or more expensive measurements are missing by design. Data can also be missing completely at random (MCAR) and unrelated to available covariates, when the test samples are predicted under cross-validation schemes in supervised machine learning, such as CART (Breiman et al. 1984) and boosting the nearest neighbor classifiers (Breiman 1996). In many empirical studies, missing data mechanisms can be analyzed as functions of relevant covariates termed missing-at-random (MAR; Rubin 1976), otherwise, missing data are generated by special causes hence termed missing-not-at-random (MNAR). Aside from informative MNAR cases such as censored data or selection-biased samples (Marlin et al. 2007), it is easy to make but nontrivial to test the MAR assumption (Qu and Song 2002; Potthoff et al. 2006) compared with MCAR (Fuchs

Corresponding author: Michelle Liou, Email adress: mliou@stat.sinica.edu.tw

1982; Diggle 1989; Chen and Little 1999). Nevertheless, an estimable MAR model is usually assumed such that inference can be carried out using available covariates, instead of deleting the incomplete units in the data.

While parametric inference for the mean of an incomplete variable is commonly examined using the EM algorithm under the MAR assumption, various methods of predicting the unidentified units has gained popularity in the machine learning literature. Instead of assuming parametric models for the regression function or missing data pattern, nonparametric regression methods have been discussed since 1980s (Matloff 1981; Cheng and Wei 1986; Altman 1992). Under MAR, asymptotic normality of the kernel regression (KR) imputation was initially examined by Cheng and Wei (1986) and Cheng (1990, 1994). A review of the existing nonparametric estimators is given for the motivation of this study. Suppose that a random sample with incomplete responses are observed and denoted as

$$(X_i, Y_i, \delta_i), i = 1, 2, \dots, n. \quad (1.1)$$

All the covariates X_i are observed, and $\delta_i = 1$ if Y_i is observed, $\delta_i = 0$ otherwise. The parameter of interest is the mean of Y ($\mu = EY$), which can be estimated under the MAR assumption, that is, the probability of missing Y depends solely on the covariate X

$$P(\delta = 1|X, Y) = P(\delta = 1|X) \equiv p(X). \quad (1.2)$$

Let $m(x) = E(Y|X = x)$ denote the regression function. Two KR imputation estimators for the mean are defined as

$$\tilde{\mu}_{KR} = \frac{1}{n} \sum_{i=1}^n \hat{m}_{KR}(X_i) \quad (1.3)$$

and

$$\hat{\mu}_{KR} = \frac{1}{n} \sum_{i=1}^n \{\delta_i Y_i + (1 - \delta_i) \hat{m}_{KR}(X_i)\}; \quad (1.4)$$

where

$$\hat{m}_{KR}(X_i) = \frac{\sum_{j=1}^n W_h(X_i, X_j) \delta_j Y_j}{\sum_{j=1}^n W_h(X_i, X_j) \delta_j}, \quad (1.5)$$

$W_h(u, x) = h^{-1}W((u - x)/h)$, W is a symmetric probability (kernel) density function in the domain of the variable X , and h is the kernel bandwidth. These two KR imputation

estimators approximate to the same normal distribution and are termed asymptotically equivalent in distribution (Cheng 1994). In the literature, the KR imputation estimator (1.4) was also discussed with semiparametric regression analysis and empirical likelihood inference (e.g., Wang et al. 2004).

A well-known alternative to the KR estimation is the k -nearest neighbor (k -NN) regression estimation. The k -NN regression has traditionally been a popular method in the machine learning literature in engineering science (Cover and Hart 1967; Toussaint 2005). The one nearest neighbor (1-NN) imputation was applied to nonresponses in survey sampling (Sande 1979), and discussed by Lee et al. (1994), Rancourt (1999), Chen and Shao (2000) and Shao and Wang (2008). The k -NN regression and imputation was discussed by Cheng (1984, 1994), and by Ning and Cheng (2012) for the prediction of the Iris species as an alternative method to the popular CART (Loh and Shih, 1997) and support vector machine (SVM; Gunn 1998). With a positive integer k , the k -NN imputation estimator for the mean is defined as

$$\hat{\mu}_{kNN} = \frac{1}{n} \sum_{i=1}^n \{\delta_i Y_i + (1 - \delta_i) \hat{m}_{kNN}(X_i)\}. \quad (1.6)$$

Here, the kernel imputation estimates $\hat{m}_{KR}(X_i)$ of (1.5) are replaced by the nearest neighbor estimates $\hat{m}_{kNN}(X_i) = \frac{1}{k} \sum_{j=1}^k Y_{i(j)}$, using the k nearest complete pairs $\{(X_{i(j)}, Y_{i(j)}) : \delta_{i(j)} = 1, j = 1, \dots, k\}$, where $X_{i(j)}$ denotes the j th nearest neighbor of X_i among the observed pairs. The fixed kernel bandwidth h of (1.5) is replaced by a random distance from X_i to its k th nearest neighbor $X_{i(k)}$ having $\delta_{i(k)} = 1$, where the Euclidean or the Mahalanobis distance may be used. Such distance functions can also be used with the KR estimator (1.5) when the covariate X is multivariate.

Another nonparametric estimator of the mean is derived from classical inverse probability weighting (IPW) due to Horvitz and Thompson (1952). It estimates the population mean using IPW to reflect the effective sample size (Cochran 1977). Under MAR, the naive Horvitz-Thompson (HT) estimator for μ is

$$\hat{\mu}_{HT} = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i Y_i}{w_i}, \quad (1.7)$$

where for each $i = 1, \dots, n$,

$$w_i \equiv \hat{p}(X_i) = \frac{\sum_{j=1}^n \delta_j W_h(X_i, X_j)}{\sum_{j=1}^n W_h(X_i, X_j)} \quad (1.8)$$

is a locally-weighted kernel estimate of the missing pattern function value $p(X_i)$, as an analog of the regression estimate (1.5). The IPW imputation estimator for the mean can be derived from the KR imputation, that is,

$$\hat{\mu}_{IPW} = \frac{1}{n} \sum_{i=1}^n \left[\hat{m}_{KR}(X_i) + \frac{\delta_i [Y_i - \hat{m}_{KR}(X_i)]}{w_i} \right], \quad (1.9)$$

where $\hat{m}_{KR}(X_i)$ and w_i are given in (1.5) and (1.8), respectively. Estimator (1.9) was obtained by replacing $m(X_i)$ and $p(X_i)$ with $\hat{m}_{KR}(X_i)$ and w_i , respectively, in the proof for the asymptotic normality of the KR estimator (1.4) (Cheng 1994, Appendix, pp. 86-87). The IPW estimator is expected to reduce the sample bias of the KR estimator at the cost of slightly increased variance, and the advantage is that the IPW often yields smaller mean squared errors (MSE). A recent simulation study showed that under regularity conditions (cf. Appendix), HT, IPW and KR yield comparable performances of the sample variances, MSE, and the coverage probabilities of confidence intervals (CCI) (cf. Ning and Cheng 2012). In theory, it can be shown that estimators (1.4) (or (1.9)) and (1.7) approximate the same normal distribution as stated in Lemma 1.1 below.

Lemma 1.1 *Imputation estimators $\hat{\mu}_{KR}$, $\hat{\mu}_{HT}$ and $\hat{\mu}_{IPW}$ approximate the same normal distribution $N(\mu, \sigma_{KR}^2)$ under a common set of regularity conditions. The asymptotic variance is denoted by*

$$\sigma_{KR}^2 = \text{Var}(Y) + E \left[\frac{\sigma^2(X)(1 - p(X))}{p(X)} \right], \quad (1.10)$$

where $\sigma^2(X) = \text{Var}(Y|X)$.

It is well known that the k -NN and kernel-weighted estimators yield smaller sample bias but larger variance when using small k or bandwidth h , and the converse occurs with larger k or h , which are the trade-off effects between sample biases and variances in choosing k or h . It is also found that when the missing pattern function has jump discontinuities or decreases towards zero over an interval in the domain of the covariate X (i.e., two typical non-regularity conditions), the sample bias and variance of

imputation are enlarged when using the KR, HT and IPW, but the k -NN is less affected by definition and yields small bias and MSE by using small k (Ning and Cheng 2012).

These facts lead to the consideration of a convex mixture of the KR estimator (1.4) and the k -NN estimator (1.6). The proposed convex mixtures (CM) imputation estimator and its IPW version (CMIPW) will be defined and examined in Section 2. In addition, a mixed combination of the CM and the IPW can be formulated as the third convex mixture imputation, termed the convex regression (CR) imputation estimator. It is proved that the CR yields smaller asymptotic variance than the k -NN under regularity conditions such that CR is expected to yield satisfactory performance under general conditions. Section 3 presents a simulation study to demonstrate improved performances of the proposed CM, CMIPW and CR estimators against the existing estimators under two typical non-regular missing data conditions. A simulation study under standard regularity conditions and another under an extremely non-regular condition are given in the Supplements. In Section 4, applications of the CM estimator to predicting unknown classes are examined using the well-known Iris species and wine-quality taste preference data sets from the UCI Machine Learning Repository (Asuncion and Newman 2007). The CM estimator acquires comparable performances to a few supervised classification methods, but it is less competitive to the SVM with complex multivariate data, because it is not designed as a supervised learning method. Section 5 concludes the study with a brief discussion on potential application of the proposed CM and CR imputation methods to general missing data and classification environments. The Appendix presents the required regularity conditions and the proofs for Lemma 1.1, Theorems 2.1 and 2.2. Two additional simulation cases of regular and non-regular missing data patterns, basic descriptive statistics of the wine quality data and related computations are given in the Supplement.

2. Convex Mixtures Imputation

In this section, new imputation methods using convex mixtures of the KR and k -NN estimators are introduced. The basic convex imputation estimator for the mean of the response variable $\mu = EY$ is

$$\hat{\mu}_{CM} = \frac{1}{n} \sum_{i=1}^n [\delta_i Y_i + (1 - \delta_i) \hat{m}_{CM}(X_i)], \quad (2.1)$$

where

$$\hat{m}_{CM}(X_i) = w_i \hat{m}_{KR}(X_i) + (1 - w_i) \hat{m}_{kNN}(X_i), \quad (2.2)$$

and w_i is given in (1.8). By (1.5) and (1.8), the first summand of the convex estimate (2.2) is a local kernel regression estimate based on the observed responses, and the second summand furnishes the k -NN regression estimate using the non-observation (missing) weight. Thus, the convex mixtures (CM) estimator (2.1) is defined using the local convex regression estimate (2.2) to balance the trade-off effect between the sample bias and variance given by the two regression estimates. Similar to reducing the bias (hence the MSWE) of the KR (1.4) by using the IPW (1.9), the IPW version of the CM estimator is defined as

$$\hat{\mu}_{CMIPW} = \frac{1}{n} \sum_{i=1}^n \left[\hat{m}_{CM}(X_i) + \frac{\delta_i [Y_i - \hat{m}_{CM}(X_i)]}{w_i} \right]. \quad (2.3)$$

By analogy with the asymptotic equivalence between KR and IPW of Lemma 1.1, the same asymptotic normality is acquired by both CM and CMIPW as stated below, and proved in the Appendix.

Theorem 2.1 *Assume the regularity conditions (H), (S) and (W) in the Appendix. The imputation estimators $\hat{\mu}_{CM}$ of (2.1) and $\hat{\mu}_{CMIPW}$ of (2.3) approximates the same normal distribution*

$$\sqrt{n}(\hat{\mu}_{CM} - \mu) \rightarrow N(0, \sigma_{CM}^2),$$

where

$$\begin{aligned} \sigma_{CM}^2 &= \text{Var}[m(X)] + E \left[\frac{\sigma^2(X)}{p(X)} \right] \\ &\quad + \frac{1}{k} E \left[\sigma^2(X) \{1 - p(X)\}^3 \left(1 + \frac{1}{k} \right) \right], \end{aligned} \quad (2.4)$$

and the first two terms on the right-hand side of (2.4) yields σ_{KR}^2 of (1.10).

The asymptotic variances of the imputation estimators, the k -NN (1.6), IPW (1.9) (or, the KR and HT), CM (2.1) and CMIPW (2.3) can be compared as follows. It was known from Ning and Cheng (2012, Theorem 1) that

$$\sigma_{kNN}^2 = \sigma_{KR}^2 + \frac{1}{k} E \left[\sigma^2(X) \{1 - p(X)\} \right]. \quad (2.5)$$

The asymptotic variances of (1.10) and (2.4) are related by the equation

$$\sigma_{CM}^2 = \sigma_{KR}^2 + \frac{1}{k} E \left[\sigma^2(X) \{1 - p(X)\}^3 \left(1 + \frac{1}{k}\right) \right]. \quad (2.6)$$

Like k -NN, under regularity conditions, the CM and CMIPW yield larger variances than KR and IPW, hence larger MSE, because these estimators are all asymptotically unbiased under regularity conditions. However, under non-regular conditions, KR and IPW may yield larger bias and MSE as mentioned after Lemma 1.1. In view of the two IPW versions, the IPW (of the KR) and CMIPW, it is possible to form a third IPW version using a combination of these two. This new version is termed a convex regression (CR) imputation estimator for the mean μ , defined as

$$\hat{\mu}_{CR} = \frac{1}{n} \sum_{i=1}^n \left[\hat{m}_{CM}(X_i) + \frac{\delta_i \{Y_i - \hat{m}_{KR}(X_i)\}}{w_i} \right]. \quad (2.7)$$

Under regularity conditions, the CR yields a different asymptotic normality from the previous estimators.

Theorem 2.2 *Under the same conditions of Theorem 2.1, the CR imputation estimator $\hat{\mu}_{CR}$ approximates the normal distribution*

$$\sqrt{n}(\hat{\mu}_{CR} - \mu) \rightarrow N(0, \sigma_{CR}^2)$$

where

$$\sigma_{CR}^2 = \sigma_{KR}^2 + \frac{1}{k} E \left[\sigma^2(X) \{1 - p(X)\}^2 \right] \quad (2.8)$$

Like σ_{CM}^2 , it is seen from (2.8) that σ_{CR}^2 is larger than σ_{KR}^2 , but σ_{CR}^2 is smaller than σ_{kNN}^2 ,

$$\sigma_{kNN}^2 = \sigma_{CR}^2 + \frac{1}{k} E \left[p(X) \{1 - p(X)\} \sigma^2(X) \right]. \quad (2.9)$$

It is understood that valid theoretical results can only be acquired under regularity conditions, that is, no theory can be derived under non-regularity conditions, because sample bias and MSE can vary widely when regularity conditions are violated. As mentioned, typical non-regular conditions include the case having jump discontinuities of the missing pattern function $p(x)$ or the conditional variance function $\sigma^2(x)$, and the case where $p(x)$ may decrease towards zero over an interval within the domain of X . To

understand the effect of non-regularity, we will examine the sampling bias, variance, MSE and the CCI of all imputation estimators under these conditions.

3. Simulation of Mean Imputation

The simulation study was designed to examine the performance of the proposed imputation estimators CM, CMIPW and CR, compared with the existing estimators, k -NN, KR, HT and IPW. Because all estimators are expected to perform almost equally well under regularity conditions (cf. Ning and Cheng, 2012), the simulation study will be conducted under two typical non-regular conditions, and those under the regularity and another conditions will be given in the Supplement. The basic form of the regression model is

$$Y = m(X) + \varepsilon, \quad (3.1)$$

where the error variable $\varepsilon \sim N(0, \sigma^2(x))$ is assumed to be independent of the covariate X . In each simulation case, the distribution of X and missing pattern function $p(x)$ were defined, and random samples of sizes $n = 100, 500, 1000$ were generated using model (3.1). Average imputation estimates of the mean EY were computed using 1,000 replications, and performances were evaluated using averaged sample bias, variance, MSE, and CCI. The report of each simulation case consists of one table, respectively. We followed two basic rules in computing the kernel estimates.

1. A common kernel function was used for all KR type estimators, that is, the Epanechnikov quadratic kernel function

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

which was used throughout the entire simulation study.

2. In case $\sum_{j=1}^n W_h(X_i, X_j)\delta_j = 0$, then $w_i = \hat{p}(X_i) = 0$, there was no candidate donor within one-bandwidth distance from the covariate X_i ; and, no values would be imputed for the missing response when using KR and IPW, and the actual sample size (reduced by one for each case) for estimating the mean could be less than n . This rule was not applied to the proposed CM, CMIPW and CR estimators because a weighted k -NN estimate $\hat{m}_{kNN}(X_i)$ was imputed by definition.

3.1 Case 1

$$\left\{ \begin{array}{l} m_1(x) = 10 - 6((x_1 + x_2) - 1.2)^2, \\ (X_1, X_2) \sim U([0, 1]^2), \\ E(Y) = 8.760, E(Y_{obs}) = 8.190, \\ P(\delta = 1) = 0.398, \sigma_{KR}^2 = 5.637, \\ \sigma_{CM}^2 = 5.637 + 0.329 \times \frac{1}{k}(1 + \frac{1}{k}), \\ \sigma_{CR}^2 = 5.637 + 0.413 \times \frac{1}{k}, \end{array} \right. \quad p_1(x) = \left\{ \begin{array}{ll} 0.7, & 0 \leq x_1 + x_2 \leq 0.6, \\ 0.2, & 0.6 < x_1 + x_2 \leq 1.4, \\ 0.8, & 1.4 < x_1 + x_2 \leq 2.0, \\ 0.16, & 0 \leq x_1 + x_2 \leq 0.6, \\ 1.0, & 0.6 < x_1 + x_2 \leq 1.4, \\ 0.16, & 1.4 < x_1 + x_2 \leq 2.0. \end{array} \right.$$

[Table 1 about here.]

In this case, jump discontinuity (line segments in a plane or points in an interval) of the missing pattern function $p(x)$ and the conditional variance function $\sigma^2(x)$ violates the regularity conditions (cf. Appendix), and no estimator can perform well with a small sample size such as 100. Being sensitive to discontinuity, both KR and HT are expected to yield large sample biases and MSEs, hence are omitted from discussion in this case, although KR always yields smaller sample variances than the IPW. In fact, large biases of the KR are indicated by the large variations of the second summand (in the definition of the IPW) in the calculated replicates of the IPW. The sample variances and MSEs of the CM are fairly stable across values of h for each k , and the average values over h are comparable to those of the k -NN for each k . Thus, the choice $k = 4$ is recommended to be used for both k -NN and CM because the average sample variance at $k = 4$ is near to the median value, but not the minimum, among a few k neighbors, due to the trade-off effect between the sample bias and variance (cf. Table 1). By analogy, for the IPW, the choice of h , corresponding to the median value among those having small sample variances when h varies from 0.15 to 0.25, is recommended (cf. Table 1).

For the proposed estimators CM and CMIPW, by analogy with the pair (KR, IPW), it is seen that CM yields smaller sample variances but larger biases versus CMIPW. An exception is that, unlike the other estimators, the sample variances of CMIPW do not proportionally decrease along with the increased sample sizes from 500 to 1000. Therefore, it is seen that, with k larger than 4, the sample MSEs of CMIPW are smaller than those of CM, k -NN and IPW when n is around 500, but the opposite holds when CM uses k (2 to 4) and IPW uses small h (less than 0.30) when n is as large as 1000. By analogy, CMIPW also yields slightly smaller sample variances and

performs better than the CR when using k (4 to 8) with sample size about 500. However, for larger sample sizes around 1000, CR gives the smallest MSEs by using the pairs $(h \in (0.15, 0.20), k = 8)$ and $(h \in (0.20, 0.25), k = 16)$ subject to the trade-off effect. In addition, it is worth noting that the choices of k and h corresponding to the smallest sample variances can be recommended for all imputation estimators under the regularity conditions, i.e., simulation Case 3, because the magnitudes of sample biases of all estimators are negligibly small such that slight differences in the sample MSEs are also negligible (cf. Tables S1 and S2).

3.2 Case 2

$$\left\{ \begin{array}{ll} m_2(x) = 2x + 1, \quad x \in (-3, 4), & E(Y) = 2.90, \quad E(Y_{obs}) = 4.740, \\ X \sim 0.3U(-3, 0) + 0.7U(0, 4), & P(\delta = 1) = 0.647, \quad \sigma_{KR}^2 = 18.004, \\ p_2(x) = \frac{e^x}{1+e^x}, & \sigma_{CM}^2 = 18.004 + 0.169 \times \frac{1}{k} \left(1 + \frac{1}{k}\right), \\ \sigma_2^2(x) = 1, & \sigma_{CR}^2 = 18.004 + 0.224 \times \frac{1}{k}. \end{array} \right.$$

[Table 2 about here.]

In this case, all regularity conditions are satisfied except that the smooth $p(x)$ decreases towards 0 only at the left-end point ($x = -3.0$) in the domain of X . It is seen that no estimator can be consistent with small sample size $n = 100$; and, KR also fails to be consistent with enlarged sample sizes. The k -NN is consistent with small k , 1 and 2, when $n = 500$; and, 1 to 4, when $n = 1000$. The performances of CM and k -NN are comparable, CM yields slightly smaller sample variances, but larger sample biases such that the choices $k = 1$ when $n = 500$, and $k = 2$ when $n = 1000$ are recommended for both estimators. It is remarkable that IPW yields larger magnitudes of sample biases than 0.04 when $n = 500$, and it is consistent only with $h \in (0.6, 1.0)$ when $n = 1000$. Indeed, both k -NN and CM perform better than the IPW by using small $k = 1$ with moderately large sample size $n = 500$.

Meanwhile, CMIPW (using $k = 1$ to 4) and CR (using $k = 1$ to 2) also perform better than the IPW when $n = 500$. It is remarkable that CMIPW yields smaller sample variances and MSEs compared with CR, although CR yields smaller sample biases. It is especially remarkable that CMIPW gives smaller sample MSEs than the IPW, when using the pairs $(h \in (0.6, 1.0), k = 4)$ with $n = 1000$. The results in this case seem to convey an important and useful message, in that the widely discussed non- and semi-

parametric IPW estimators may not perform well at moderate sample sizes about 500, and it does not yield the smallest MSEs with larger sample sizes about 1000 when the popular logistic missing pattern function decreases towards 0 at the domain boundary of the covariate X .

In the Supplement, simulation Case 3 (Tables S1 and S2) presents essentially equal performances of all the estimators under regularity conditions. Simulation Case 4 (Table S3) presents the same model conditions of Case 2, but replacing the exponent x of $p(x)$ in Case 2 with $2.5x$ such that $p(x)$ decreases to 0 at a very fast rate at an end point of the covariate X . In this case, it is found that only CR is able to yield consistent imputation by giving sample biases of magnitudes less than 0.10 and the CCI estimates greater than 0.90 when using the bandwidths $h \in (0.6, 1.0)$, and when the sample sizes are larger than 2000.

4. Empirical Study of Prediction Effect

4.1 Iris Flower Data

The iris flower data set has been much discussed on the classification of the iris species since the linear discrimination analysis of Fisher (1936). The data set consists of 50 samples from each of three species of iris flowers: setosa, virginica and versicolor. Four features were measured for each sampled species, the descriptions and scatter plots can be found in the UCI Machine Learning Repository (MLR, <http://archive.ics.uci.edu/ml>). The attribute of interest is the species classification variable, denoted by Y , where $Y = 0$ defines the iris setosa, $Y = 1$ the virginica and $Y = 2$ the versicolor. The useful covariates are the four-dimensional predictor vectors $X = (X_1, X_2, X_3, X_4)$, i.e., the natural features (Sepal length, Sepal width, Petal length, Petal width). By the two-variable scatter plots, the species may be well classified using a discrimination rule such as CART based on the joint distribution of $(X_3 = \text{petal length}, X_4 = \text{petal width})$ without the need of using other features, for example, Loh and Shih (1997). Suppose that regression prediction methods are evaluated using training and test samples of the iris data, for which an MAR design is used to define observed (training) and missing (test) samples. Without using structural relations between the species and the features in the bivariate data plots as in CART, the least informative evenly-spread feature X_2 , the sepal width, can be used to generate observed and missing samples. As in Ning and Cheng (2012), the evaluation was based on

simulating a fixed missing pattern in the entire training-and-testing procedure. That is,

$$p(x) = \begin{cases} 0.7, & x_2 < 0.3 \\ 0.1, & x_2 \geq 0.3, \end{cases} \quad (4.1)$$

where $E[p(X)] = 0.328$, and about two-thirds of the species types were generated as missing in the evaluation study. To predict the unobserved iris species Y in discrete responses, kernel imputation estimates of (1.5) may be modified as

$$\hat{m}_{KR}(X_i) = \arg \max_t \{a_{KR}(X_i, t)\},$$

where

$$a_{KR}(X_i, t) = \frac{\sum_{j=1}^n \delta_j W_h(X_i, X_j) I(Y_j = t)}{\sum_{j=1}^n W_h(X_i, X_j) \delta_j}, \quad t = 0, 1, 2 \quad (4.2)$$

and $I(Y_j = t) = 1$ when $Y_j = t$; otherwise, it is equal to 0. The k -NN imputation estimates are modified from (1.6) as

$$\hat{m}_{NN}(X_i) = \arg \max_t \{a_{NN}(X_i, t)\},$$

where

$$a_{NN}(X_i, t) = \frac{1}{k} \sum_{j=1}^k I(Y_{i(j)} = t), \quad t = 0, 1, 2. \quad (4.3)$$

Similarly, the proposed CM imputation estimates (2.2) are modified as

$$\hat{m}_{CM}(X_i) = \arg \max_t \{a_{CM}(X_i, t)\},$$

where

$$a_{CM}(X_i, t) = \hat{p}(X_i) a_{KR}(X_i, t) + (1 - \hat{p}(X_i)) a_{NN}(X_i, t), \quad (4.4)$$

for $t = 0, 1, 2$ and the weight estimates $\hat{p}(X_i)$ are defined in (1.8). For each replicated data set, imputation estimates (4.2) to (4.4) were used to predict the unknown (missing) species using the observed training sample. The simulation was repeated five hundred times ($n = 500$), the average prediction accuracy (PA)

$$PA = \frac{\sum_{i=1}^n (1 - \delta_i) I(\hat{m}(X_i) = Y_i)}{\sum_{i=1}^n (1 - \delta_i)} \quad (4.5)$$

and its standard error (se) were calculated for the predicted values $\hat{m}(X_i) = \hat{m}_{KR}(X_i)$, $\hat{m}_{NN}(X_i)$ and $\hat{m}_{CM}(X_i)$ of (4.2) to (4.4), as listed in Tables 3 and 4 below.

[Table 3 about here.]

[Table 4 about here.]

In Table 3, the best choice of k ($k = 1$) for the k -NN is seen to yield the minimal sample standard error of prediction (0.0148) with the maximal average PA (0.9697), or 3.03% prediction error rate. Similarly, the KR yields its maximal average PA (0.9609) using a bandwidth about $h = 1.2$ by its minimal sample standard error (0.0191). In Table 4, the proposed CM imputation method outperforms the k -NN and KR by using the pair ($h \in (0.6, 1.0)$, $k = 1$ or 2) with minimal sample standard error (0.013 - 0.015). That is, CM yields the best average PA (0.970 to 0.971), or 3.0% average prediction error rate. A 95% confidence interval of the PA statistics of the CM prediction ($k = 2$) is given as 0.970 ± 0.025 , or (0.945, 0.995), which is slightly preferred to the CART results of Loh and Shih (1997), and the SVM results of Gunn (1998).

4.2 Wine Quality Data

Cortez et al. (2009) proposed a method based on the SVM to predict human wine taste preference using a data set of Portugal white and red vinho verde wine samples, which consists of 1599 red wine samples and 4898 white wine samples. For both samples, 11 physicochemical features denoted as X_1, \dots, X_{11} are recorded; and the output variable, denoted by Y , is the wine quality graded score between 0 (poor) and 10 (excellent). Supplementary Table S4 lists the descriptive statistics of these features of both samples according to Cortez et al. (2009, Table 1). In their study, the authors trained variable selection by SVM and analyzed the variability of the wine quality Y when the chosen predictor features vary with different levels while holding the other features fixed at the average levels. The predictor variables causing the largest variability on Y were regarded as the most relevant features to the wine quality. For both red and white wine data, these 11 predictor features were ranked. The most important features related to the red wine quality were found as pH (X_9), sulphates (X_{10}) and total sulfur dioxide (X_7). For white wine, they were sulphates (X_{10}), alcohol (X_{11}) and residual sugar (X_4).

Similar to the study of the iris data, we borrowed a probabilistic missing data mechanism approach to evaluating the prediction of unknown (artificial missing) wine quality scores under a convenient MAR design. For the current multi-dimensional features of wine quality, three most relevant features to the wine quality were used to

define missing data patterns for red and white wine data samples, respectively,

$$P_{red}(\text{unobserved}|x) = p_{red}(x) = \frac{1}{1 + e^{(x_7+x_9+x_{10})}} \quad (4.6)$$

$$P_{white}(\text{unobserved}|x) = p_{white}(x) = \frac{1}{1 + e^{(x_4+x_{10}+x_{11})}} \quad (4.7)$$

The average missing rate for the red wine was 0.5104 by equation (4.6), and that for the white wine was 0.4781 by (4.7). In contrast, the average missing rates (proportions of test samples by the cross-validation design) for both red and white wine were about 0.33 in the study of Cortez et al (2009). It is known that missing rates closer to 0.50 are statistically more fairly-judged than 0.33 in the evaluation of cross-validation. Missing patterns (4.6) and (4.7) were repeatedly simulated 500 times for each wine data, where the predicted regression estimates of the missing units were calculated using the observed (training) sample. Following Wang et al (2003) and Cortez et al (2009), two evaluation criteria for the prediction performance were measured: the mean absolute deviation (MAD)

$$MAD = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (4.8)$$

and the PA within a tolerance limit T ($T = 0.25, 0.50$ and 1.0)

$$PA(T) = \frac{1}{n} \sum_{i=1}^n 1_{[0,T]}(|y_i - \hat{y}_i|). \quad (4.9)$$

Here, y_i denotes the unobserved true integer score, \hat{y}_i denotes the predicted value for y_i , n is the number of unobserved units and $1_A(\cdot)$ is the indicator function. The average MAD and corresponding PA(T) values of the three predictors, the KR, k -NN and CM, and their standard errors were calculated from these 500 simulated replicates, respectively. The bandwidth parameter and the number k of nearest neighbors could be directly selected according to the best average performance among the training samples. Estimates of the MAD and corresponding PA(T) values, and their standard errors in predicting the missed (tested) wine sample integer scores Y were recorded in Table 5. For a comparison study, we also list in the same tables the performance results of three existing classifiers, multiple linear regression (MR), neural network (NN, cf. Sun et al. 1997) and SVM as given in Cortez et al. (2009, Table 2). Modified

calculations using unequal k -NN weights can be found in Tables S5 to S8. The minimal average MAD, the corresponding PA(T) values and standard errors are listed against the parameter values h , k and (h, k) on the left-hand side column of each table. It is notable that weighted k -NN estimators (using unequal k -NN weights) were used with the CM predictor, because uniform weights ($1/k$) could inflate the PA values when using a closed tolerance interval $[0, T]$ in the definition (4.9) (cf. Cortez et al. 2009, Section 2.2). To predict integer raw scores, such inflation of estimated PA values could arise when $T = 0.50$ or 1.0 , and when moderately small k (2, 4, 8 and 16) were used with the k -NN and CM predictors, which could yield estimates having decimals exactly equal to 0.50 or 0.0. As shown in Table 5, the performances of both KR and k -NN, being plain regression predictors, are not expected to be as equally effective as the existing methods. Improving upon the KR and k -NN prediction, the proposed CM predictor yields comparable performance to the classical MR and NN, giving slightly greater PA values at the standard tolerance level $T = 0.50$. The results also show that the CM is inferior to the SVM due to two obvious facts. First, the CM was used for testing about 50% missing units in the data by the missing patterns (4.6) and (4.7), whereas the SVM was used for testing one-third missing units (Cortez et al. 2009). Second, the missing patterns (4.6) and (4.7) were not defined or modified by a sensitivity analysis between the 11 features and the incomplete response variable. That is, it differs from the main-stream supervised learning methods by using the same (h, k) pair of parameters in the CM method with both training and testing samples throughout the entire simulation process without supervised learning or updated modification. This explains why the CM prediction costs the least computational effort when compared with the supervised learning methods, MR, NN and SVM.

[Table 5 about here.]

5. Discussion

For survey data with missing items of individual units, missing data patterns can be estimated using the basic KR regression estimates as functions of continuous covariates without assuming parametric models under the MAR condition. It is shown that the proposed convex mixtures of regression imputation estimators for the mean, the CM, CMIPW and CR, may yield more stable or improved performance over the k -NN, KR and IPW under non-regular missing pattern functions and joint distributions,

which accommodate general missing data conditions in practice. However, it is cautioned in Section 3 that all imputation estimators, but the CR, can fail to be consistent if the observation rates $p(x)$ may decrease rapidly towards zero in the domain of the predictor X .

In the machine learning literature, the k -NN regression, multiple linear regression and neural networks have been widely used in classification (e.g., Huang, et al. 2004; Kiang 2003). By using families of kernel functions with optimization techniques, the SVM is able to yield more efficient classification (Smola and Scholkopf 2004). In the empirical study of the Iris data in Section 4, the proposed CM prediction is shown to yield improved performance over the KR and k -NN, and improved confidence interval over both CART and SVM. For the wine quality data with more sophisticated joint distribution, the CM prediction is found to yield comparable performance to the classical MR and NN, but less satisfactory results when compared with the SVM. This is so, because CM is only a semi-supervised learning method, which uses the least amount of computation among these methods. In a future study, it is expected to improve the performance of the CM prediction when the simple missing mechanisms (4.6) and (4.7) are replaced by functions that will carry more information between the response variable and the explanatory features in the wine quality data. Meanwhile, the estimators CMIPW and CR are potentially useful for providing multiple imputation estimates for possible improvement of prediction accuracy.

In conclusion, we remark that the introduction of the convex mixture imputation using the CM prediction, the CMIPW and CR estimation can be useful with general missing data in practice. For example, the topic of adaptive local estimation of the regression function under general missing data conditions is a basic problem for which the proposed estimators are applicable. In the world of data mining, it is remarkable that the proposed CM can often yield improved prediction over the k -NN, but further study is needed to enhance the prediction accuracy of the CM by using more data information.

Acknowledgment We thank the reviewers for their useful comments on an early version of the manuscript. This research was supported by grants to Ning from NSFC No. 11571133, No. 11471135 and China Scholarship Council, and to Liou and Cheng from MOST No. 103-2410-H-001-058-MY2.

Appendix

The proofs for Lemma 1.1 and Theorems 2.1 and 2.2 will need a few regularity conditions on the kernel, the bandwidth, the regression function $m(x)$ and the missing pattern $p(x)$ as listed below.

- (W) The kernel function W is a symmetric probability density function (pdf) defined on a bounded interval on the line or d -dimensional Euclidean R^d such that $\int |u|^2 W(u) du$ is finite.
- (H) The kernel weights $W_h(u, x) = h^{-1} W((u - x)/h)$ are defined with a decreasing sequence of bandwidths $h (= h(n))$ such that $h(n) \rightarrow 0$, $nh^2 \rightarrow \infty$ and $nh^4 \rightarrow 0$. Extensions of condition (h) to the R^d case can be found in Cheng (1994, Appendix).
- (S) EY^2 and $E[\sigma^2(X)/p(X)]$ are finite. The regression function $m(x)$, the missing pattern function $p(x)$ and the conditional variance function $\sigma^2(X)$ have bounded second-order derivatives, and $p(x)$ cannot be decreasing towards zero in an interval within the domain of the covariate X .

Proof of Lemma 1.1

We will first prove that the KR imputation estimator (1.4) and the IPW imputation estimator (1.9) are asymptotically equivalent in distribution, that is, both $\sqrt{n}(\hat{\mu}_{KR} - \mu)$ and $\sqrt{n}(\hat{\mu}_{IPW} - \mu)$ approximate the same normal distribution (1.10) as $n \rightarrow \infty$. Then, a proof for the same property of the HT estimator $\hat{\mu}_{HT}$ (1.7) will follow immediately.

First, let us recall the proof of asymptotic normality for $\hat{\mu}_{KR}$ (Cheng, 1994, Appendix). There, a basic expression is $\hat{\mu}_{KR} - \mu = R + S + T_{KR}$, where $R = \frac{1}{n} \sum_{i=1}^n [m(X_i) - \mu]$, $S = \frac{1}{n} \sum_{i=1}^n \delta_i (Y_i - m(X_i))$, and $T_{KR} = \frac{1}{n} \sum_{i=1}^n (1 - \delta_i) (\hat{m}_{KR}(X_i) - m(X_i))$. It was shown that T_{KR} is asymptotically equivalent to U_n in distribution (i.e., $\sqrt{n}(T_{KR} - U_n) \rightarrow 0$ in probability as $n \rightarrow \infty$) by mean-square convergence of $\hat{m}_{KR}(X)$ to $m(X)$. This is expressed by the formula

$$T_{KR} \approx U_n = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i (Y_i - m(X_i))(1 - p(X_i))}{p(X_i)} \quad (\text{A.1})$$

Omitting the parameter μ , it follows by (A.1) and formula (1.9) that

$$\begin{aligned} \hat{\mu}_{KR} &\simeq \frac{1}{n} \sum_{i=1}^n m(X_i) + S + \frac{1}{n} \sum_{i=1}^n \frac{\delta_i(Y_i - m(X_i))(1 - p(X_i))}{p(X_i)} \\ &= \frac{1}{n} \sum_{i=1}^n m(X_i) + \frac{1}{n} \sum_{i=1}^n \frac{\delta_i(Y_i - m(X_i))}{p(X_i)} \\ &\simeq \hat{\mu}_{IPW} \end{aligned} \tag{A.2}$$

In the last step of (A.2), by analogy with the proof for (A.1), the asymptotic equivalence to $\hat{\mu}_{IPW}$ is justified when $m(X_i)$ and $p(X_i)$ are estimated (and replaced) by $\hat{m}_{KR}(X_i)$ of (1.5) and w_i of (1.8), respectively. This proves that $\hat{\mu}_{IPW}$ approximates $N(\mu, \sigma_{IPW}^2)$ and $\sigma_{IPW}^2 = \sigma_{KR}^2$ as in (1.10). It is worth noting that the same arguments for (A.1) and (A.2) also prove that the naive kernel estimator (1.3) approximates the same normal distribution (Cheng and Wei, 1986).

Next, we will show that $\hat{\mu}_{HT}$ also approximates the same normal distribution in the first step above. Writing $\hat{\mu}_{HT} - \mu = R + S_{HT}$, where $S_{HT} = n^{-1} \sum_{i=1}^n \left(\frac{\delta_i Y_i}{w_i} - m(X_i) \right)$, the first fact is that the conditional expectation of S_{HT} given the covariates $\{X_i, i = 1, 2, \dots, nx\}$ is asymptotically negligible, or of magnitude $o(1/\sqrt{n})$ in probability. This follows by checking that

$$\begin{aligned} E \left[E \left(\frac{\delta_i Y_i}{w_i} | X_i \right) - m(X_i) \right] &= E \left[E \left(\frac{p(X_i)}{w_i} - 1 \right) m(X_i) \right] \\ &= E \left[\frac{\sum_j W_h(X_j, X_i) \{p(X_j) - p(X_i)\}}{\sum_j W_h(X_j, X_i) \delta_j} m(X_i) \right] \\ &= O(h^2) \end{aligned} \tag{A.3}$$

which yields the desired result because $\sqrt{nh^2} \rightarrow 0$ by condition (h). It also implies that the variance of the sample average of left-hand side of (A.3) is asymptotically negligible. Because $Var(R) = Var[m(X)]$, it remains to show that the expectation of the conditional variance of S_{HT} given the covariates $\{X_i\}$'s yields the desired approximate variance. This can be expressed as

$$\begin{aligned} E \left[Var \left(\frac{\delta_i Y_i}{w_i} | X_i \right) \right] &= E \left[\frac{p(X_i) \sigma^2(X_i)}{p^2(X_i) (1 + O(h^2))} \right] \\ &= E \left[\frac{\sigma^2(X_i)}{p(X_i) (1 + O(h^2))} \right] \\ &\simeq E \left[\frac{\sigma^2(X)}{p(X)} \right] \end{aligned} \tag{A.4}$$

The sum of $Var(R)$ and the right-hand-side of (A.4) yields the desired asymptotic variance $\sigma_{HT}^2 (= \sigma_{KR}^2)$. This concludes the proof for Lemma 1.1.

Proof of Theorem 2.1

As in the proof of Lemma 1.1, the imputation estimator (2.1) can be expressed as

$$\hat{\mu}_{CM} = \frac{1}{n} \sum_{i=1}^n m(X_i) + S + T_{CM}, \quad (\text{A.5})$$

where $T_{CM} = \frac{1}{n} \sum_{i=1}^n (1 - \delta_i)(\hat{m}_{CM}(X_i) - m(X_i))$, and the convex imputation estimate is $\hat{m}_{CM}(X_i) = w_i \hat{m}_{KR}(X_i) + (1 - w_i) \hat{m}_{kNN}(X_i)$ of (2.2). It follows from T_{KR} of (A.1) that the sample average of the first summand $w_i \hat{m}_{KR}$ is asymptotically equivalent to the term T_1 of (A.6) below. And, the average of the second summand $(1 - w_i) \hat{m}_{kNN}$ is asymptotically equivalent to the last summand T_2 in (A.6), which follows from the proof for the asymptotic normality of $\hat{\mu}_{kNN}$ (Ning and Cheng (2012, Appendix)). These two facts are put together to yield

$$\begin{aligned} T_{CM} &= \frac{1}{n} \sum_{i=1}^n (1 - \delta_i) [w_i \{\hat{m}_{KR}(X_i) - m(X_i)\} \\ &\quad + (1 - w_i) \{\hat{m}_{kNN}(X_i) - m(X_i)\}] \\ &\simeq \frac{1}{n} \sum_{i=1}^n \delta_i (Y_i - m(X_i)) \{1 - p(X_i)\} (\equiv T_1) \\ &\quad + \frac{1}{n} \sum_{i=1}^n (1 - \delta_i) \{1 - p(X_i)\} \left[\frac{1}{k} \sum_{j=1}^k (Y_{i(j)} - m(X_{i(j)})) \right] (\equiv T_2). \end{aligned} \quad (\text{A.6})$$

By (A.6), it is seen that (A.5) is asymptotically equivalent to

$$\begin{aligned} \hat{\mu}_{cm} - \mu &= \frac{1}{n} \sum_{i=1}^n \{m(X_i) - \mu\} (\equiv R) \\ &\quad + \frac{1}{n} \sum_{i=1}^n \delta_i (Y_i - m(X_i)) (2 - p(X_i)) (\equiv S + T_1) + T_2 \\ &= R + \frac{1}{n} \sum_{i=1}^n \delta_i (Y_i - m(X_i)) (2 - p(X_i)) (\equiv S^*) + T_2. \end{aligned} \quad (\text{A.7})$$

It follows from (A.7) that $E(R) = E(S^*) = E(T_2) = 0$, $Cov(R, S^*) = 0 = Cov(R, T_2)$, $nVar(R) = Var[m(X)]$,

$n\text{Var}(S^*) = E[p(X)\{2 - p(X)\}^2\sigma^2(X)]$ and the remaining covariance terms are

$$\begin{aligned} n\text{Var}(T_2) &= \frac{1}{k}E\left[(1 - p(X))^3\sigma^2(X)\right] + E\left[\frac{(1 - p(X))^3\sigma^2(X)}{p(X)}\right]; \\ 2n\text{Cov}(S^*, T_2) &= 2E\left[(1 - p(X))^2\{2 - p(X)\}\sigma^2(X)\right]. \end{aligned} \quad (\text{A.8})$$

The sum of the above three variance terms and the covariance term in (A.8) is

$$\begin{aligned} \sigma_{CM}^2 &= \text{Var}[m(X)] + E\left[\frac{\sigma^2(X)}{p(X)}\right] + \frac{1}{k}E\left[\sigma^2(X)\{1 - p(X)\}^3\left(1 + \frac{1}{k}\right)\right] \\ &= \sigma_{KR}^2 + \frac{1}{k}E\left[\sigma^2(X)\{1 - p(X)\}^3\left(1 + \frac{1}{k}\right)\right]. \end{aligned} \quad (\text{A.9})$$

The right-hand side of (A.9) is equal to (2.4), which proves Theorem 2.1.

Proof of Theorem 2.2

By definition, the convex imputation estimator (2.7) is expressed as

$$\begin{aligned} \hat{\mu}_{CR} - \mu &= \frac{1}{n} \sum_{i=1}^n \left[\hat{m}_{CM}(X_i) + \frac{\delta_i\{Y_i - \hat{m}_{KR}(X_i)\}}{w_i} \right] - \mu \\ &= R + \frac{1}{n} \sum_{i=1}^n w_i \{\hat{m}_{KR}(X_i) - m(X_i)\} \\ &\quad + \frac{1}{n} \sum_{i=1}^n (1 - w_i) \{\hat{m}_{kNN}(X_i) - m(X_i)\} + \frac{1}{n} \sum_{i=1}^n \frac{\delta_i\{Y_i - \hat{m}_{KR}(X_i)\}}{w_i} \end{aligned}$$

which can be shown, by analogy with (A.6), as asymptotically equivalent to

$$\hat{\mu}_{CR} - \mu \simeq R + S + \frac{1}{n} \sum_{i=1}^n \{1 - p(X_i)\} \left[\frac{1}{k} \sum_{i=1}^k \{Y_{i(j)} - m(X_{i(j)})\} \right] \quad (\text{A.10})$$

using the same R and S of (A.1). It follows by a similar analysis to (A.8) that the variance of (A.10) is equal to

$$\begin{aligned} \sigma_{CR}^2 &= \text{Var}[m(X)] + E[p(X)\sigma^2(X)] + 2E[\{1 - p(X)\}\sigma^2(X)] \\ &\quad + \frac{1}{k}E\left[\sigma^2(X)\{1 - p(X)\}^2\right] + E\left[\frac{\sigma^2(X)\{1 - p(X)\}^2}{p(X)}\right] \\ &= \text{Var}[m(X)] + E\left[\frac{\sigma^2(X)}{p(X)}\right] + \frac{1}{k}E\left[\sigma^2(X)\{1 - p(X)\}^2\right]. \end{aligned} \quad (\text{A.11})$$

The right-hand side of (A.11) is equal to (2.8), and the proof for Theorem 2.2 is complete.

References

- Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, **46**, 175-185.
- Asuncion, A. and Newman, D. (2007). UCI Machine Learning Repository, University of California, Irvine, <http://www.ics.uci.edu/mlearn/MLRepository.html>.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, **24**, 123-140.
- Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*. New York: Chapman and Hall.
- Chen, H. Y. and Little, R. A. (1999). A test of missing completely at random for generalized estimating equations with missing data. *Biometrika*, **86**, 1-13.
- Chen, J. and Shao, J. (2000). Nearest neighbor imputation for survey data. *J. official Stat.*, **16**, 113-132.
- Chen, J. and Shao, J. (2001). Jackknife variance estimation for nearest neighbor imputation. *J. Amer. Stat. Assoc.* **96**, 260-269.
- Cheng, P. E. (1984). Strong consistency of nearest neighbor regression function estimators. *J. Multiv. Anal.*, **15**, 63-72.
- Cheng, P. E. (1990). Applications of kernel regression estimation: A survey. *Commun. Statist.-Theory Meth.*, **19**, 4103-4134.
- Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *J. Amer. Stat. Assoc.*, **89**, 81-87.
- Cheng, P. E. and Chu, C. K. (1996). Kernel estimation of distribution functions and quantiles with missing data. *Statistica Sinica*, **6**, 63-78.
- Cheng, P. E. and Wei, L. J. (1986). Nonparametric inference under ignorable missing data process and treatment assignment. *International Statistical Symposium*, Taipei, **1**, 97-112.

- Cochran, W. G. (1977). *Sampling Techniques*. New York: Wiley.
- Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *In Decision Support Systems*, **47**, 547-553.
- Cover, T. M. and Hart, P. E. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, **13**, 21-27.
- Diggle, P. J. (1989). Testing for random dropouts in repeated measurement data. *Biometrics*, **45**, 1255-1258.
- Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. *J. Amer. Stat. Assoc.*, **77**, 270-278.
- Gunn, S. R. (1998). Support vector machines for classification and regression. Technical Report MP-TR-98-05, Image Speech and Intelligent Systems Group, University of Southampton.
- Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite population. *J. Amer. Stat. Assoc.*, **47**, 663-685.
- Huang, Z., Chen, H., Hsu, C., Chen, W. and Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, **37**, 543-558.
- Kiang, M. (2003). A comparative assessment of classification methods. *Decision Support Systems*, **35**, 441-454.
- Lee, H., Rancout, E. and Sarndal, C. E. (1994). Experiments with variance estimation from survey data with imputed values. *J. Official Stat.*, **10**, 231-243.
- Loh, W. Y. and Shih, Y. S. (1997). Split selection methods for classification trees. *Statist. Sinica*, **7**, 815-840.
- Marlin, B. M., Zemel, R. S., Rowels, S. and Slaney, M. (2007). Collaborative filtering and the missing at random assumption. *In Proceedings of the 23rd conference on Uncertainty in Artificial Intelligence*.

- Matloff, N. S. (1981). Use of regression functions for improved estimation of means. *Biometrika*, **68**, 685-689.
- Ning, J. H. and Cheng, P. E. (2012). A comparison study of nonparametric imputation methods. *Stat. Comput.*, **22**, 273-285.
- Potthoff, R. F., Tudor, G. E., Pieper, K. S. and Hasselblad, V. (2006). Can one assess whether missing data are missing at random in medical studies? *Statistical Methods in medical Research*, **15**, 213-234.
- Qu, A. and Song, P. (2002). Testing ignorable missingness in estimating equation approaches for longitudinal data. *Biometrika*, **89**, 841-850.
- Rancourt, E. (1999). Estimation with nearest neighbor imputation at Statistics Canada. in *Proceedings of the Section on Survey Research Methods*, Amer. Statist. Assoc., 131-138.
- Rubin, D. B. (1976). Inference and missing data, *Biometrika*, **63**, 581-592.
- Sande, I. G. (1979). A personal view of Hot Deck imputation procedures. *Survey Methodology*, **5**, 238-258.
- Shao, J. and Wang, H. (2008). Confidence intervals based on survey data with nearest neighbor imputation. *Stat. Sinica*, **18**, 281-297.
- Smola, A. and Scholkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, **14**, 199-222.
- Sun, L., Danzer, K. and Thiel, G. (1997). Classification of wine samples by means of artificial neural networks and discrimination analytical methods. *Fresenius' Journal of Analytical Chemistry*, **359**, 143-149.
- Toussaint, G. T. (2005). Geometric proximity graphs for improving nearest neighbor methods in instance-based learning and data mining. *International Journal of Computational Geometry and Applications*, **15**, 101-150.
- Wang, Q., Linton, O. and Hardle, W. (2004). Semiparametric regression analysis with missing response at random. *J. Amer. Statist. Assoc.*, **99**, 334-345.

Wang, W., Xu, Z., Lu, W. and Zhang, X. (2003). Determination of the spread parameter in the Gaussian kernel for classification and regression. *Neurocomputing*, **55**, 643-663.

School of Mathematics and Statistics, Central China Normal University, 152 Luoyu Road, Wuhan, Hubei, P. R. China 430079

E-mail: jhning@mail.ccnu.edu.cn

Institute of Statistical Science, Academia Sinica, 128 Academia Road, Section 2, Nankang, Taipei 115, Taiwan

E-mail: mliou@stat.sinica.edu.tw

Institute of Statistical Science, Academia Sinica, 128 Academia Road, Section 2, Nankang, Taipei 115, Taiwan

E-mail: pcheng@stat.sinica.edu.tw

Table 1: Average estimates using different imputation methods under simulation Case 1

| Method | | | Sample Size n | | | | | | | | | | | |
|---------|------|----|-----------------|----------------------|----------------------|-------|--------|----------------------|----------------------|-------|--------|----------------------|----------------------|-------|
| | | | 100 | | | | 500 | | | | 1000 | | | |
| h | k | | Bias | $n \cdot \text{VAR}$ | $n \cdot \text{MSE}$ | CCI | Bias | $n \cdot \text{VAR}$ | $n \cdot \text{MSE}$ | CCI | Bias | $n \cdot \text{VAR}$ | $n \cdot \text{MSE}$ | CCI |
| KR | 0.10 | - | -0.244 | 8.473 | 14.436 | 0.877 | -0.036 | 6.239 | 6.893 | 0.936 | -0.019 | 5.672 | 6.014 | 0.937 |
| | 0.15 | - | -0.134 | 7.044 | 8.828 | 0.924 | -0.035 | 5.625 | 6.233 | 0.937 | -0.035 | 5.341 | 6.569 | 0.917 |
| | 0.20 | - | -0.093 | 6.230 | 7.103 | 0.937 | -0.059 | 5.426 | 7.146 | 0.919 | -0.060 | 5.196 | 8.756 | 0.873 |
| IPW | 0.15 | - | -0.126 | 7.115 | 8.704 | 0.928 | -0.013 | 5.988 | 6.072 | 0.952 | -0.011 | 5.639 | 5.751 | 0.934 |
| | 0.20 | - | -0.072 | 6.355 | 6.879 | 0.942 | -0.017 | 5.771 | 5.917 | 0.942 | -0.016 | 5.480 | 5.723 | 0.929 |
| | 0.25 | - | -0.057 | 6.035 | 6.363 | 0.945 | -0.024 | 5.579 | 5.873 | 0.941 | -0.023 | 5.306 | 5.818 | 0.924 |
| k -NN | - | 2 | -0.078 | 5.954 | 6.560 | 0.936 | -0.016 | 5.937 | 6.072 | 0.950 | -0.009 | 5.933 | 6.016 | 0.949 |
| | - | 4 | -0.134 | 5.859 | 7.644 | 0.916 | -0.029 | 5.522 | 5.944 | 0.943 | -0.015 | 5.583 | 5.819 | 0.935 |
| | - | 8 | -0.236 | 6.219 | 11.804 | 0.864 | -0.055 | 5.487 | 7.008 | 0.924 | -0.028 | 5.357 | 6.130 | 0.923 |
| CM | 0.10 | 2 | -0.075 | 5.964 | 6.528 | 0.936 | -0.017 | 5.908 | 6.051 | 0.952 | -0.011 | 5.851 | 5.967 | 0.943 |
| | 0.15 | 2 | -0.076 | 5.934 | 6.508 | 0.937 | -0.020 | 5.838 | 6.040 | 0.948 | -0.015 | 5.766 | 5.976 | 0.941 |
| | 0.10 | 4 | -0.125 | 5.884 | 7.448 | 0.916 | -0.028 | 5.552 | 5.937 | 0.944 | -0.016 | 5.555 | 5.807 | 0.937 |
| | 0.15 | 4 | -0.124 | 5.844 | 7.371 | 0.918 | -0.031 | 5.522 | 5.989 | 0.944 | -0.019 | 5.507 | 5.882 | 0.935 |
| | 0.10 | 8 | -0.217 | 6.180 | 10.880 | 0.871 | -0.050 | 5.478 | 6.718 | 0.930 | -0.026 | 5.356 | 6.045 | 0.925 |
| | 0.15 | 8 | -0.211 | 6.102 | 10.559 | 0.871 | -0.052 | 5.455 | 6.816 | 0.927 | -0.029 | 5.332 | 6.195 | 0.924 |
| CMIPW | 0.20 | 4 | -0.098 | 6.387 | 7.344 | 0.932 | -0.017 | 5.838 | 5.984 | 0.953 | -0.009 | 6.051 | 6.137 | 0.956 |
| | 0.25 | 4 | -0.089 | 6.303 | 7.088 | 0.931 | -0.019 | 5.662 | 5.848 | 0.953 | -0.012 | 5.937 | 6.071 | 0.956 |
| | 0.30 | 4 | -0.086 | 6.190 | 6.932 | 0.931 | -0.023 | 5.532 | 5.788 | 0.953 | -0.015 | 5.853 | 6.072 | 0.955 |
| | 0.20 | 8 | -0.132 | 6.451 | 8.194 | 0.924 | -0.019 | 5.571 | 5.755 | 0.951 | -0.012 | 5.921 | 6.064 | 0.955 |
| | 0.25 | 8 | -0.111 | 6.290 | 7.527 | 0.932 | -0.022 | 5.427 | 5.668 | 0.949 | -0.015 | 5.799 | 6.010 | 0.953 |
| | 0.30 | 8 | -0.103 | 6.121 | 7.175 | 0.934 | -0.026 | 5.311 | 5.649 | 0.949 | -0.018 | 5.721 | 6.045 | 0.950 |
| | 0.20 | 16 | -0.186 | 7.035 | 10.506 | 0.901 | -0.025 | 5.522 | 5.828 | 0.947 | -0.015 | 5.781 | 5.998 | 0.947 |
| | 0.25 | 16 | -0.154 | 6.665 | 9.034 | 0.917 | -0.028 | 5.391 | 5.775 | 0.945 | -0.018 | 5.665 | 5.995 | 0.947 |
| | 0.30 | 16 | -0.141 | 6.337 | 8.319 | 0.923 | -0.033 | 5.284 | 5.820 | 0.940 | -0.022 | 5.592 | 6.092 | 0.945 |
| CR | 0.10 | 4 | -0.124 | 5.910 | 7.449 | 0.917 | -0.021 | 5.784 | 6.009 | 0.951 | -0.007 | 5.826 | 5.876 | 0.944 |
| | 0.15 | 4 | -0.118 | 5.889 | 7.272 | 0.922 | -0.010 | 5.869 | 5.918 | 0.955 | 0.004 | 5.885 | 5.900 | 0.944 |
| | 0.20 | 4 | -0.106 | 5.884 | 7.012 | 0.929 | 0.005 | 5.907 | 5.918 | 0.949 | 0.018 | 5.899 | 6.219 | 0.941 |
| | 0.10 | 8 | -0.216 | 6.203 | 10.859 | 0.870 | -0.043 | 5.666 | 6.603 | 0.930 | -0.017 | 5.548 | 5.850 | 0.932 |
| | 0.15 | 8 | -0.205 | 6.129 | 10.337 | 0.873 | -0.032 | 5.706 | 6.202 | 0.944 | -0.006 | 5.590 | 5.629 | 0.942 |
| | 0.20 | 8 | -0.192 | 6.072 | 9.746 | 0.885 | -0.016 | 5.732 | 5.868 | 0.945 | 0.008 | 5.636 | 5.702 | 0.944 |
| | 0.20 | 16 | -0.302 | 7.071 | 16.177 | 0.812 | -0.058 | 5.596 | 7.261 | 0.923 | -0.012 | 5.448 | 5.594 | 0.934 |
| | 0.25 | 16 | -0.281 | 6.842 | 14.753 | 0.827 | -0.040 | 5.547 | 6.336 | 0.931 | 0.005 | 5.389 | 5.412 | 0.942 |
| | 0.30 | 16 | -0.260 | 6.697 | 13.445 | 0.844 | -0.021 | 5.505 | 5.735 | 0.946 | 0.022 | 5.337 | 5.825 | 0.935 |

Table 2: Average estimates using different imputation methods under simulation Case 2

| Method | | Sample Size n | | | | | | | | | | | | |
|---------|-----|-----------------|----------|----------|--------|-------|----------|----------|--------|-------|----------|----------|--------|-------|
| | | 100 | | | | 500 | | | | 1000 | | | | |
| h | k | Bias | n -VAR | n -MSE | CCI | Bias | n -VAR | n -MSE | CCI | Bias | n -VAR | n -MSE | CCI | |
| KR | 0.4 | - | 0.535 | 32.754 | 61.344 | 0.835 | 0.092 | 24.756 | 28.995 | 0.929 | 0.034 | 19.490 | 20.665 | 0.943 |
| | 0.6 | - | 0.401 | 33.764 | 49.840 | 0.886 | 0.081 | 20.308 | 23.563 | 0.932 | 0.052 | 17.641 | 20.342 | 0.925 |
| | 0.8 | - | 0.331 | 32.020 | 42.986 | 0.897 | 0.100 | 17.961 | 22.921 | 0.917 | 0.083 | 16.972 | 23.864 | 0.905 |
| IPW | 0.6 | - | 0.384 | 33.853 | 48.610 | 0.891 | 0.049 | 20.769 | 21.955 | 0.941 | 0.015 | 18.148 | 18.381 | 0.950 |
| | 0.8 | - | 0.297 | 32.208 | 41.055 | 0.904 | 0.041 | 18.676 | 19.512 | 0.949 | 0.019 | 17.582 | 17.936 | 0.948 |
| | 1.0 | - | 0.244 | 29.812 | 35.751 | 0.921 | 0.044 | 17.821 | 18.795 | 0.950 | 0.026 | 17.455 | 18.153 | 0.945 |
| k -NN | - | 1 | 0.116 | 21.753 | 23.104 | 0.953 | 0.027 | 18.101 | 18.477 | 0.947 | 0.009 | 18.847 | 18.935 | 0.950 |
| | - | 2 | 0.184 | 21.142 | 24.509 | 0.939 | 0.042 | 18.044 | 18.914 | 0.947 | 0.014 | 18.079 | 18.285 | 0.948 |
| | - | 4 | 0.309 | 20.637 | 30.179 | 0.906 | 0.069 | 18.179 | 20.582 | 0.940 | 0.027 | 18.064 | 18.766 | 0.951 |
| CM | 0.2 | 1 | 0.117 | 21.653 | 23.013 | 0.950 | 0.028 | 17.997 | 18.380 | 0.951 | 0.010 | 18.587 | 18.697 | 0.948 |
| | 0.4 | 1 | 0.119 | 21.558 | 22.965 | 0.949 | 0.031 | 17.950 | 18.430 | 0.950 | 0.014 | 18.534 | 18.726 | 0.947 |
| | 0.2 | 2 | 0.180 | 21.127 | 24.360 | 0.940 | 0.042 | 17.973 | 18.839 | 0.948 | 0.015 | 18.003 | 18.234 | 0.946 |
| | 0.4 | 2 | 0.181 | 21.060 | 24.337 | 0.938 | 0.045 | 17.944 | 18.948 | 0.948 | 0.019 | 17.975 | 18.320 | 0.947 |
| | 0.2 | 4 | 0.295 | 20.674 | 29.386 | 0.912 | 0.068 | 18.087 | 20.380 | 0.939 | 0.027 | 17.981 | 18.691 | 0.952 |
| | 0.4 | 4 | 0.294 | 20.622 | 29.263 | 0.911 | 0.071 | 18.069 | 20.573 | 0.939 | 0.030 | 17.967 | 18.866 | 0.946 |
| CMIPW | 0.6 | 1 | 0.116 | 21.528 | 22.884 | 0.952 | 0.027 | 17.935 | 18.312 | 0.951 | 0.010 | 18.508 | 18.607 | 0.948 |
| | 0.8 | 1 | 0.117 | 21.512 | 22.879 | 0.951 | 0.028 | 17.906 | 18.304 | 0.949 | 0.011 | 18.480 | 18.594 | 0.946 |
| | 1.0 | 1 | 0.118 | 21.497 | 22.890 | 0.951 | 0.030 | 17.877 | 18.316 | 0.949 | 0.012 | 18.457 | 18.605 | 0.946 |
| | 0.6 | 2 | 0.147 | 21.586 | 23.749 | 0.947 | 0.030 | 17.890 | 18.342 | 0.952 | 0.012 | 18.099 | 18.233 | 0.947 |
| | 0.8 | 2 | 0.141 | 21.745 | 23.726 | 0.948 | 0.031 | 17.849 | 18.336 | 0.951 | 0.013 | 18.014 | 18.175 | 0.945 |
| | 1.0 | 2 | 0.138 | 21.803 | 23.698 | 0.951 | 0.034 | 17.730 | 18.300 | 0.953 | 0.014 | 17.983 | 18.192 | 0.945 |
| | 0.6 | 4 | 0.201 | 21.806 | 25.845 | 0.938 | 0.035 | 17.865 | 18.465 | 0.953 | 0.013 | 17.685 | 17.865 | 0.953 |
| | 0.8 | 4 | 0.184 | 22.098 | 25.485 | 0.941 | 0.036 | 17.790 | 18.428 | 0.952 | 0.015 | 17.618 | 17.851 | 0.950 |
| | 1.0 | 4 | 0.175 | 22.210 | 25.271 | 0.938 | 0.039 | 17.601 | 18.381 | 0.948 | 0.018 | 17.621 | 17.938 | 0.950 |
| CR | 0.2 | 1 | 0.116 | 21.705 | 23.042 | 0.953 | 0.024 | 18.120 | 18.416 | 0.949 | 0.007 | 18.790 | 18.838 | 0.950 |
| | 0.4 | 1 | 0.112 | 21.635 | 22.881 | 0.951 | 0.016 | 18.259 | 18.386 | 0.952 | -0.005 | 19.062 | 19.085 | 0.952 |
| | 0.6 | 1 | 0.104 | 21.693 | 22.773 | 0.954 | -0.001 | 18.491 | 18.491 | 0.947 | -0.023 | 19.111 | 19.630 | 0.948 |
| | 0.2 | 2 | 0.179 | 21.156 | 24.352 | 0.940 | 0.038 | 18.021 | 18.753 | 0.947 | 0.012 | 18.114 | 18.250 | 0.948 |
| | 0.4 | 2 | 0.174 | 21.106 | 24.135 | 0.939 | 0.030 | 18.153 | 18.597 | 0.948 | -0.000 | 18.396 | 18.396 | 0.954 |
| | 0.6 | 2 | 0.166 | 21.173 | 23.927 | 0.941 | 0.013 | 18.340 | 18.427 | 0.950 | -0.018 | 18.480 | 18.808 | 0.949 |
| | 0.2 | 4 | 0.294 | 20.697 | 29.348 | 0.907 | 0.064 | 18.097 | 20.171 | 0.939 | 0.023 | 18.038 | 18.573 | 0.952 |
| | 0.4 | 4 | 0.287 | 20.652 | 28.885 | 0.915 | 0.056 | 18.200 | 19.754 | 0.944 | 0.011 | 18.273 | 18.402 | 0.953 |
| | 0.6 | 4 | 0.278 | 20.725 | 28.443 | 0.919 | 0.039 | 18.360 | 19.121 | 0.953 | -0.007 | 18.339 | 18.385 | 0.955 |

Table 3: Average KR and k-NN prediction accuracy for the iris species

| KR | | <i>k</i> -NN | |
|-----|-----------------|--------------|-----------------|
| h | PA (s.e.) | k | PA (s.e.) |
| 0.8 | 0.9312 (0.0380) | 1 | 0.9697 (0.0148) |
| 1.0 | 0.9488 (0.0262) | 2 | 0.9628 (0.0198) |
| 1.2 | 0.9609 (0.0191) | 4 | 0.9597 (0.0439) |
| 1.5 | 0.9502 (0.0196) | 8 | 0.9220 (0.1413) |
| 1.8 | 0.9370 (0.0228) | - | - |

*PA (s.e.) values are averages of 500 replicates

Table 4: Average CM prediction accuracy for the iris species

| $h \setminus k$ | 1 | 2 | 4 | 8 |
|-----------------|---------------|---------------|---------------|---------------|
| 0.1 | 0.970 (0.015) | 0.966 (0.014) | 0.960 (0.044) | 0.956 (0.070) |
| 0.2 | 0.970 (0.015) | 0.966 (0.014) | 0.961 (0.043) | 0.955 (0.070) |
| 0.5 | 0.970 (0.015) | 0.970 (0.015) | 0.965 (0.044) | 0.937 (0.102) |
| 0.6 | 0.970 (0.015) | 0.971 (0.014) | 0.966 (0.044) | 0.932 (0.116) |
| 0.8 | 0.970 (0.015) | 0.970 (0.013) | 0.965 (0.043) | 0.927 (0.131) |
| 1.0 | 0.970 (0.015) | 0.970 (0.013) | 0.965 (0.043) | 0.925 (0.138) |
| 1.2 | 0.970 (0.015) | 0.967 (0.015) | 0.961 (0.043) | 0.922 (0.140) |
| 1.5 | 0.970 (0.015) | 0.964 (0.017) | 0.957 (0.044) | 0.918 (0.141) |
| 1.8 | 0.970 (0.015) | 0.963 (0.018) | 0.957 (0.044) | 0.918 (0.141) |

*PA (s.e.) values are averages of 500 replicates

Table 5: Average MAD and PA(T) values for the red and white wine data

| Wine | Method | MAD | PA (T = 0.50) | PA (T = 1.0) |
|------------|--------------------------------|----------------------|----------------------|----------------------|
| White Wine | KR (s.e.) (h = 2.5) | 0.551 (0.011) | 0.568 (0.016) | 0.880 (0.009) |
| | MR (C.I.) | 0.500 (\pm 0.000) | 0.591 (\pm 0.001) | 0.886 (\pm 0.001) |
| | k-NN (s.e.) (k = 32) | 0.537 (0.011) | 0.569 (0.016) | 0.879 (0.009) |
| | NN (C.I.) | 0.51 (\pm 0.000) | 0.591 (\pm 0.003) | 0.888 (\pm 0.002) |
| | CM (s.e.) (h = 0.5, k = 32) | 0.514 (0.011) | 0.596 (0.016) | 0.886 (0.009) |
| | SVM (C.I.) | 0.46 (\pm 0.000) | 0.624 (\pm 0.004) | 0.890 (\pm 0.002) |
| Red Wine | KR (s.e.) (h = 2.5) | 0.617 (0.009) | 0.490 (0.009) | 0.832 (0.006) |
| | MR (C.I.) | 0.59 (\pm 0.000) | 0.517 (\pm 0.001) | 0.843 (\pm 0.001) |
| | k-NN (s.e.) (k=32) | 0.600 (0.009) | 0.515 (0.011) | 0.830 (0.006) |
| | NN (C.I.) | 0.58 (\pm 0.000) | 0.526 (\pm 0.003) | 0.847 (\pm 0.001) |
| | CM (s.e.) (h = 0.5, k = 32) | 0.583 (0.009) | 0.534 (0.011) | 0.837 (0.006) |
| | SVM (C.I.) | 0.45 (\pm 0.000) | 0.646 (\pm 0.004) | 0.868 (\pm 0.004) |

*PA(T) values using different $h, k, (h, k)$ are given in Tables S5-S8