

# Kernel-Based Multi-imputation for Missing Data\*

Shichao ZHANG<sup>1</sup>, Yongsong QIN<sup>2</sup>, Xiaofeng ZHU<sup>2</sup>, Jilian ZHANG<sup>2</sup>, Chengqi ZHANG<sup>1</sup>

*1 Faculty of Information Technology, University of Technology Sydney*

*2 Department of Computer Science, Guangxi Normal University, China*

**ABSTRACT:** A Kernel-Based Nonparametric Multiple imputation method is proposed under MAR (Missing at Random) and MCAR (Missing Completely at Random) missing mechanisms in nonparametric regression settings. We experimentally evaluate our approach, and demonstrate that our imputation performs better than the well-known NORM algorithm.

**Keywords:** multiple imputation, missing data, kernel function, nonparametric.

## 1. Introduction

Missing or incomplete data is a very important problem in many fields of research, Such as in active media technology, opinion polls, market research surveys, mail enquiries, medical studies, and other scientific experiments. Missing data imputation is a challenging issue in machine learning and data mining (Zhang et al. 2004; Batista & Monard 2003). Many missing data analysis techniques are of single-imputation which missing values are filled in by a plausible estimate such as the mean or median for that variable on other participants. However, single-imputation cannot provide valid standard errors and confidence intervals, since it ignores the uncertainty implicit in the fact that the imputed values are not the actual values. Recently, much research on missing data analysis has focused on multi-imputation techniques for addressing the issues in single-imputation (Faris et al. 2002; Little et al. 1987; Schaffer 2002; Taylor et al. 2002; Zhang 2004). Little et al. (1987) proposed a multiple imputation procedure to replace each missing value with a set of plausible values that represent the uncertainty about the right value to impute. The multiple-imputed-data sets are then analyzed using a standard procedure for complete data and combining the results from these analyses.

In this paper, a kernel-based nonparametric multiple-imputation (KBNM) is proposed under MAR (missing  $Y$  mainly depends on  $X$ ) and MCAR (MCAR is when the probability of missing a value is the same for all variables). The rest of this paper is organized as follows. Our kernel-based multiple imputation method is described in Section 2. Section 3 presents a series of experimental results on simulation models and a real-world dataset (from UCI) to compare the performances between our KBNM approach and the NORM. Conclusions are given in Section 4.

---

Corresponding author: Shichao Zhang, Faculty of Information Technology, University of Technology Sydney  
PO Box 123, Broadway NSW 2007, Australia; E-mail: zhangsc@it.uts.edu.au.

This work is partially supported by Australian large ARC grants (DP0449535, DP0559536 and DP0667060), a China NSF major research Program (60496327), and a China NSF grant (60463003).

## 2. KBNM Method

The theoretical underpinnings of multi-imputation are Bayesian. The central idea is to fill in the missing values by drawing from the posterior predictive distribution of the missing data given the observed data. The procedure is independently repeated  $M$  times. Each filled-in dataset is analyzed separately and the results combined following well-established rules. Rubin's multiple imputation is a three-step method for handling complex missing data.

At the first step,  $m$  ( $> 1$ ) completed-data sets are created by imputing the unobserved data  $m$  times using  $m$  independent draws from an imputation model, which is constructed to reasonably approximate the true distributional relationship between the unobserved data and the available information, and thus reduce potentially very serious nonrespondent bias due to systematic difference between the observed data and the unobserved ones.

In this article, we use the kernel-based nonparametric imputation to impute the nonrespondent (missing) and obtain  $m$  'complete' data sets. Let  $X$  be a  $d$ -dimensional vector of factors and  $Y$  be a respondent variable influenced by  $X$ . Suppose that  $(X_i, Y_i)$   $s$  satisfy the following model:  $Y_i = m(X_i)$ , where  $Y$  have nonrespondents and  $m(\cdot)$  is an unknown function, and  $X_i$   $s$  are i.i.d. (independence identified distributed) random variables and all  $X_i$   $s$  are observed. Let  $r = \sum_{i=1}^n \delta_i, m = n - r$  ( $n$  is sample size), denote the sets of respondents and nonrespondents as  $S_r$  and  $S_m$ , respectively. Let  $Y_i^{(R)}, i \in S_m$  be the imputed values  $Y_i^{(R)} = \hat{m}_n(X_i) + \varepsilon_i^*$ ,  $i \in S_m$ , where  $\{\varepsilon_i^*\}$  is a simple random sample of size  $m$

with replacement from  $\{Y_j - \hat{m}_n(X_j), j \in S_r\}$ .  $\hat{m}_n(x) = \frac{\sum_{i=1}^n \delta_i Y_i K(\frac{x - X_i}{h})}{\sum_{i=1}^n \delta_i K(\frac{x - X_i}{h}) + n^{-2}}$  based on the completely

observed pairs  $(X_i, Y_i)$ .

Note that:  $\delta_i = 0$  if  $Y_i$  is missing, otherwise  $\delta_i = 1$ ;  $h = h_n$  be a bandwidth sequence that decreases toward 0 as the sample size  $n$  increases toward  $\infty$ ; the term  $n^{-2}$  is introduced to avoid the case that the denominator from becoming zero.  $\kappa(\cdot)$  is a symmetric probability density function and claimed kernel function. In practice, there is no any significant difference using kernel functions and we use the Gaussian kernel (standard normal density function:  $K(\cdot) = (2\pi)^{-1/2} \exp(-x^2/2), x \sim N(1,1)$ ) in our experiments.

At the second step,  $m$  complete data analyses are performed by treating each complete data set as a real complete-data set, and thus standard complete-data procedures and software can be utilized directly.

At the last step, the results from the  $m$  complete-data analyses are combined in a simple, appropriate way to obtain the so-called repeated-imputation inference, which properly takes into account the uncertainty in the imputed values.

We use two methods analyze the  $m$  complete-data to analyze the performance of KBNM. Suppose that our primary interest lies in a scalar  $Q$  (in this article, we specify  $Q$  as the mean of the response variable); and  $m$  complete datasets under the nonparametric regression model are obtained. In our first method, we constructed a  $100(1-\alpha)\%$  interval estimate for  $Q$  based on Rubin (1987; 1999) where  $\alpha$  is the significance level and we set

$\alpha=0.05$  throughout the paper (Other values of  $\alpha$  can be chosen in practice). Another method is the RE (Relative Efficiency) which use the finite m imputation estimator, rather than using an infinite number for the fully efficient imputation, in units of variance, is approximately a function of m and  $\lambda$  (for  $\lambda$  and the formula, pleas see (Yuan 2001)).

Below Table 1 shows the relative efficiencies with different values of m and  $\lambda$  based on the formula in (Yuan 2001). For cases with little missing information, only a small number of imputations are necessary for our MI analysis, such as, the RE is 0.9662 and the repeat times is only 10 while the missing rate access to 70%. Due to lack of space, the repeat times of our experiment is 10 in the paper.

**Table 1. Relative efficiencies (RE) with different values of m and  $\lambda$**

| m  | $\lambda$ |        |        |        |        |
|----|-----------|--------|--------|--------|--------|
|    | 10%       | 20%    | 30%    | 50%    | 70%    |
| 3  | 0.9677    | 0.9375 | 0.9091 | 0.8571 | 0.8108 |
| 10 | 0.9901    | 0.9901 | 0.9852 | 0.9756 | 0.9662 |

### 3. Experiments

In order to show the effectiveness of the proposed method, extensive experiments were done on simulation models as well as real dataset using a DELL Workstation PWS650 with 2G main memory, 2.6G CPU, and WINDOWS 2000.

In our experiments, we evaluate the performances of the proposed method in making inference for the mean (Q) of the response variable. We compare the performances of the NORM (Schafer 1999) and our KBNM according to their coverage probabilities (CP) and average lengths of confidence intervals (AL) based on our constructed confidence intervals, as well as we compare the two methods with relative efficiencies (RE) according to table 1.

The NORM is a Windows 95/98/NT program for multiple imputation (MI) of incomplete multivariate data downloaded from Schafer (1999). It creates multiple imputations by an algorithm called data augmentation (DA), a special kind of Markov chain Monte Carlo (MCMC) technique. NORM is not designed to replace well-established statistical packages like SAS or SPSS and does not perform statistical analyses (e.g. linear or logistic regression).

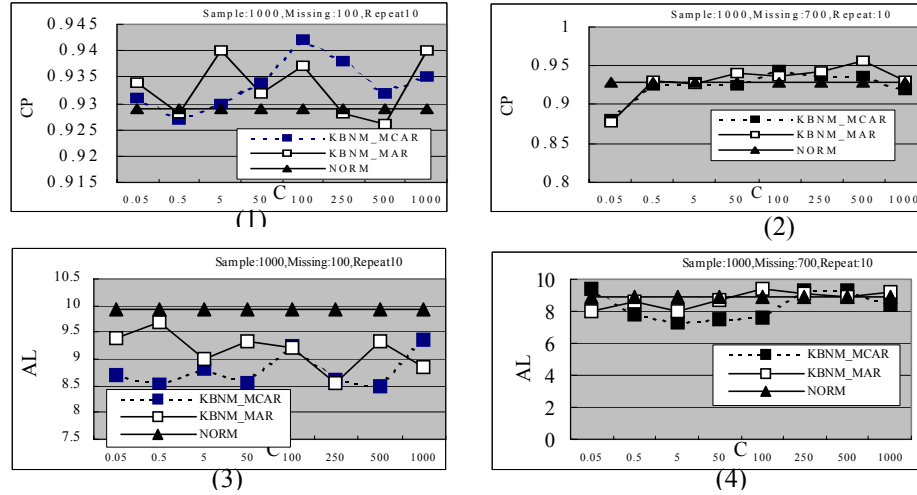
#### 3.1 Simulations

In general, we may select any nonlinear model where variables can be any attribution in simulated experiment. In this paper, we use the nonlinear model  $y=x_1^2+\sin x_2+\varepsilon$ , with  $x_i(i=1,2)$  from the normal distribution  $N(1,1)$  and  $\varepsilon$  from  $N(0, 1)$ . The following two cases of response probabilities under the MAR ( $P(\delta=1|Y,X)=P(\delta=1|X)$ ) and MCAR assumptions are considered:

*Case1* (MAR):  $P_1(x)=P(\delta=1|X=x_1, X=x_2)=0.8+0.2|x_1-1||x_2-1|$ , if  $|x_1-1||x_2-1|\leq 1$ , and =0.95, elsewhere.

*Case2* (MCAR):  $P(x) = P(\delta = 1|X=x_1, X = x_2) = 0.9$ , for all  $x_1, x_2$  respectively.

Figures (1) and (3) present the CP and AL (coverage probability and the average length of intervals) based on NORM and KBNM with various bandwidth  $h=Ct^{-1/5}$  and missing rate 10% as well as repeat times 10; Figures (2) and (4) present the same experiment as the former but the missing rate is 70%. We notice that the CP/AL of NORM is not related to the value of C, which keeps the same with various C in figures.



Figures (1) to (4) reveal the following results:  
 When the missing rate is relatively small (for example, 10%), the confidence interval based on KBNM under both of MCAR and MAR perform almost uniformly better than those based on NORM for various C as shown in Figures (1) and (3) as the CPs based on KBNM are closer to the nominal level 95% than the CP based on NORM, and the ALs are shorter based on KBNM than the AL based on NORM. For most of the C, these advantages of KBNM over the NORM are significant; When the missing rate is relatively large (for example, 70%), the confidence interval based on KBNM under both of MCAR and MAR perform still better, but not as significantly as the missing rate 10%.

By choosing appropriate C, we see that the performance of KBNM is significant better than NORM under different response rates and missing mechanisms. We would consider the choice of C in the future work.

We also compare the performances of KBNM and the NORM in terms of the RE (Relative Efficiency). Table 2 shows the Relative Efficiency in different values of m (repeat times) and  $\lambda$ , in which 10% and 70% are the missing rates.

**Table 2 The comparison of RE among KBNM under MCAR and MAR, and NORM**

| m  | $\lambda$ |        |        |        |        |        |
|----|-----------|--------|--------|--------|--------|--------|
|    | 10%       |        |        | 70%    |        |        |
|    | MCAR      | MAR    | NORM   | MCAR   | MAR    | NORM   |
| 3  | 0.9993    | 0.9948 | 0.9909 | 0.9899 | 0.9829 | 0.9550 |
| 10 | 0.9999    | 0.9994 | 0.9952 | 0.9998 | 0.9993 | 0.9792 |

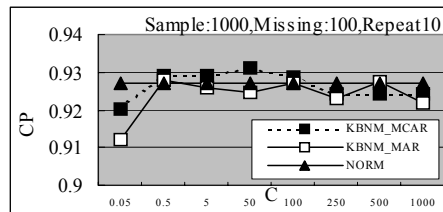
From Table 2 comparing with the standard in Table 1, we can see that all the performances of KBNM and the NORM are perform well than the performance in theory based on table 1 and the KBNM is a little better than the NORM.

### 3.2 Application in Abalone from UCI

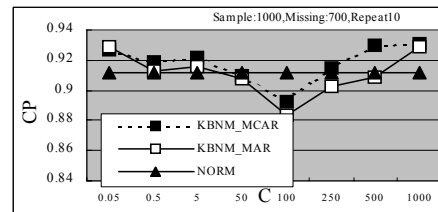
In order to show the effectiveness of our proposed method (KBNM) in making inference for the mean of a population, we conducted some experiments on the real dataset *abalone* from UCI (Blake & Merz 1998). It contains 4177 instances in total and 9 attributes for each instance, in which there are no missing values. These attributes are used to predict the age of abalone. Obviously, the relation between the age and these attributes is MAR. But we also have experiments for the data set about MACR since we want to show the difference among the three. We select the other attributes (except the “sex” who is a nominal) to predict the age of the abalone.

In this paper, we randomly select 1000 instances from 4177 because the maximum instance that NORM can only handle is 2000. We use MCAR, MAR missing mechanisms on  $Y$  at different missing rate of 10% and 70%, then the proposed nonparametric method is utilized to fill up the missing values of  $Y$ , with repeated times 10.

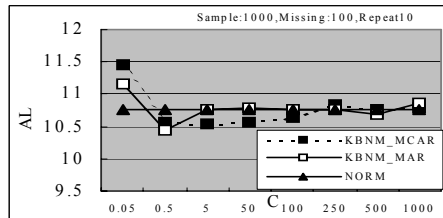
Figures (5) and (7) present the CP and AL based on NORM and KBNM with various bandwidth  $h=Cn^{-1/5}$  and missing rate 10%; Figures (6) and (8) present the performance with missing rate 70%. Table 3 shows the Relative Efficiencies in different values of  $m$  (repeat times) and  $\lambda$ . Due to the fact that the real world data do not fit the ideal statistical distributions exactly and there are noises, which will distort the distribution of the real world data, From these experiments, the performance of KBNM have a little fluctuation than the previous simulation study, but we can see that the KBNM performs better than NORM similar to the findings in appropriate C.



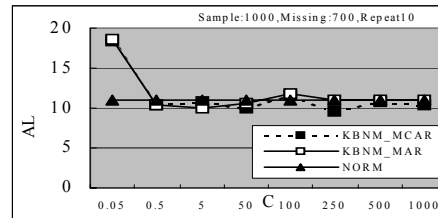
(5)



(6)



(7)



(8)

**Table 3 The comparison of RE among KBNM under MCAR and MAR, and NORM**

| m  | $\lambda$ |        |        |        |        |        |
|----|-----------|--------|--------|--------|--------|--------|
|    | 10%       |        |        | 70%    |        |        |
|    | MCAR      | MAR    | NORM   | MCAR   | MAR    | NORM   |
| 3  | 0.9983    | 0.9984 | 0.9969 | 0.9905 | 0.9926 | 0.9519 |
| 10 | 0.9997    | 0.9997 | 0.9991 | 0.9956 | 0.9959 | 0.9858 |

#### 4. Summary

We have designed an algorithm of Kernel-Based Nonparametric Multiple imputation (KBNM) to impute the incomplete datasets under MAR and MCAR assumptions. We have experimentally evaluated the performances of our KBNM and the NORM using a simulation dataset and a real dataset. The performances are in terms of the confidence intervals and the Relative Efficiencies based on different imputation methods. It has shown that our KBNM performs much better than the NORM in terms of coverage probabilities, average length of the confidence intervals and their relative efficiencies are similarly well.

#### Reference

- [1] Barnard, J.& Rubin, D. (1999). Small-Sample Degrees of Freedom with Multiple Imputation. *Biometrika*, 86: 948–955.
- [2] Batista, G. and Monard, M. (2003), An Analysis of Four Missing Data Treatment Methods for Supervised Learning. *Applied Artificial Intelligence*, 17(5-6): 519-533.
- [3] Blake, C. and Merz, C. (1998). *UCI Repository of machine learning database*. [<http://www.ics.uci.edu/~mlearn/MLResoesitory.html>] Irvine, CA: university of California, Department of Information and Computer Science.
- [4] Faris, P., et al. (2002). Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. *Journal of Clinical Epidemiology*, 55: 184–191.
- [5] Kahl, F. Heyden, A. and Quan L. (2001), Minimal Projective Reconstruction Including Missing Data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(4): 418-424.
- [6] Gessert G. (1991), Handling Missing Data by Using Stored Truth Values. *SIGMOD Record*, 20(3): 30-42.
- [7] Lakshminarayan, K., Harp, S., Goldman, R. and Samad, T. (1996), Imputation of Missing Data Using Machine Learning Techniques. *KDD-1996*: 140-145.
- [8] Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.
- [9] Schafer J. (1997). *Analysis of incomplete multivariate data*. 1st ed. London: Chapman and Hall.
- [10] Schafer, J. (1999). NORM: Multiple imputation of incomplete multivariate data under a normal model. *Version 2*. Available: <http://www.stat.psu.edu/~jls/misoftwa.html>.
- [11] Schaffer, J. (2002). Dealing with Missing Data. *Res. Lett. Inf. Math. Sci.*, 3, 153-160.
- [12] Shichao Zhang, Chengqi Zhang and Qiang Yang (2004), Information Enhancement for Data Mining. *IEEE Intelligent Systems*, 19(2): 12-13.
- [13] Taylor, J., Murray, S. & Hsu, C. (2002): Survival estimation and testing via multiple imputation. *Statistics & Probability*, 58: 221-232.
- [14] Yuan, Y.C. (2001). Multiple imputation for missing data: concepts and new development SAS/STAT 8.2. (see <http://www.sas.com/statistics>) SAS Institute Inc, Cary, NC.