Appl Intell (2007) 27:79–88
DOI 10.1007/s10489-006-0032-0

Semi-parametric optimization for missing data imputation

Yongsong Qin · Shichao Zhang · Xiaofeng Zhu · Jilian Zhang · Chengqi Zhang

Received: 15 July 2006 / Accepted: 10 November 2006 / Published online: 18 January 2007

Abstract Missing data imputation is an important issue in machine learning and data mining. In this paper, we propose a new and efficient imputation method for a kind of missing data: semi-parametric data. Our imputation method aims at making an optimal evaluation about Root Mean Square Error (RMSE), distribution function and quantile after missing-data are imputed. We evaluate our approaches using both simulated data and real data experimentally, and demonstrate that our stochastic semi-parametric regression imputation is much better than existing deterministic semi-parametric regression imputation in efficiency and effectiveness.

This work is partially supported by Australian large ARC grants (DP0449535, DP0559536 and DP0667060), a China NSF major research Program (60496327), China NSF grants (60463003, 10661003), an Overseas Outstanding Talent Research Program of Chinese Academy of Sciences (06S3011S01), a High-level Studying-Abroad Talent Program of the China Human-Resource Ministry and an Innovation Project of Guangxi Graduate Education (2006106020812M35).

S. Zhang (✉) · C. Zhang
School of Automation, Beihang University, Beijing, China
e-mail: zhangsc@it.uts.edu.au

C. Zhang
e-mail: chengqi@it.uts.edu.au

Y. Qin · X. Zhu · J. Zhang
Deparment of Computer Science, Guangxi Normal University, China
e-mail: ysqin@mailbox.gxnu.edu.cn

X. Zhu
e-mail: zhu0011@21cn.com

J. Zhang
e-mail: zhangjilian@yeah.net

Keywords Missing data · Missing data imputation · Semi-parametric data

1 Introduction

In machine learning and data mining applications [3, 11, 12], about 20% of the project effort is spent on data understanding, about 60% on data preparation and about 10% on data mining and analysis of knowledge [2]. Industrial practice also indicates that over 80% of the learning (or mining) work concentrate on data preparation [30, 31].

Indeed, data in real world applications are often missing values. Missing values may generate bias and affect the quality of the supervised learning process or the performance of classification algorithms. However, extant learning algorithms are based on the existence of quality data. In other words, researchers have been assuming that the input to the learning algorithms confirms to well-defined data distributions, containing no missing, inconsistent, or incorrect values. This leaves a large gap between the available data and the machinery available to process the data. Accordingly, this paper describes a kernel-based sem-parametric regression strategy for *missing data* imputtion.

A semi-parametric regression model is the form of $Y_i = X_i^T \beta + g(T_i) + \varepsilon_i$, where the Y_i's are i.i.d (independent identically distributed) scalar response variables, the X_i's are i.i.d d-dimensional random covariate vectors, the T_i's are i.i.d d^*-dimensional random covariate vectors, the function $g(\cdot)$ is unknown, and the model error ε_i are i.i.d random errors with mean 0 and unknown finite variance σ^2 (in our paper, we treat the unknown finite variance as 1).

Consider the sale of ice cream in summer. Generally, weather, sale place, or some unpredictable reasons can impact the sale of ice cream. Certainly, there is an important

linear relation between the sale of ice cream and weather. However, it is difficult to really know the relation between the sale of ice cream and other factors. This means that we cannot determine the real relation between the sale of ice cream and all the factors. In this paper, we construct a kernel-based stochastic semi-parametric method to handle with this complex relation. We regard it as a semi-parametric model which consists of two parts: One part is a parametric model capturing a linear model such as the relation between the sale of ice cream and weather in this example; and the other is a non-parametric model simulating such as the relation between the sale of ice cream and the other factors.

Our semi-parametric regression imputation aims to significantly overcome some shortcomings in linear models and non-parametric models by making an optimal inference on: RMSE (Root Mean Square Error), distribution function ($\theta = F(y)$) and quantile (θ_q). The distribution function $F(y)$ is the probability of Y being smaller than or equal to the given y (where y is a fixed point in R), θ_q (the q-th quantile of Y) is the level of Y that satisfies $P(Y \leq \theta_q) = q$, $0 < q < 1$. The median of Y (the case of $q = 1/2$) is the most important case of quantiles. RMSE is the accuracy of prediction.

The rest of this paper is organized as follows. We briefly outline some work related to semi-parametric imputation in Section 2. In Section 3, a kernel-based semi-parametric regression imputation is proposed in semi-parametric settings. We then use the standard statistical methods to evaluate the performance of our algorithm on RMSE, distribution function and quantile in Section 4. We summarize this paper in Section 5.

2 Related work

Missing values can be caused by error, equipment failure, change of plans, and so on. Missing values in a dataset are common in real world applications. They may lead to bias in the data, and affect the quality of learning process or the performance of knowledge discovery. Generally, methods to deal with missing data can be classified into two categories as follows: (a) Case deletion, or Learning without handling with missing data; and (b) Missing data imputation.

Case deletion, also known as listwise deletion (LD) and complete-case analysis, is the most common approach that simply omits those cases with missing values and to run analysis on only the remains. Although case deletion often results in a substantial decrease in the sample size available for the analysis, it does have important advantages. In particular, under the assumption that data are missing completely at random (MCAR, which will be introduced below), it leads to unbiased parameter estimates, and this method is suitable in the situation when the amount of missing data is small. However, if missing data are not in MAR, bias will appear which makes the results non-generalizable to the overall population. Case deletion, which gets complete data through decreasing the original data, will lose a lot of resources and information, especially, when the rate of missing data is larger or the distribution of missing data is non-random. The method can result in very serious bias and erroneous conclusion [23]. Learning with no handling with missing data, such as Bayesian Networks [16] and Artificial Neural Networks [7], is directly learning in dataset with missing value. Baysian Networks perform well when we have a prior acknowledge about the dataset or the relation among the variables in dataset are clearly understood. Otherwise, the algorithm complexity will exponentially increase due to the increasing of variables and there is an expensive cost for maintenance. Meanwhile, there are so many variables that need to be estimated, which will bring in a high variation for the Bayesian system, affecting its predicting accuracy. Furthermore, there exists the disaster of exponential explode when the dataset contains a high rate of missing rate. The technique of Artificial Neural Networks (ANN) can efficiently deal with missing values, but the research about its theory must go further.

Missing data imputation is a procedure that replaces the missing values in a dataset by some plausible values. One advantage of this approach is that the missing data treatment is independent of the learning algorithm used. This allows users to select the most suitable imputation method for their applications. However, Dempster et al. [4] pointed out: Imputation is a general and flexible method for handling with missing data problems, but is not without its pitfalls. Caution should be taken when employing imputation methods as they can generate substantial biased between real and imputed data.

There exist many techniques to manage data with missing values, but no one is absolutely better than the others. Different situations require different solutions. Allison [1] said: "the only really good solution to the missing value problem is not to have any missing in dataset". In addition, the efficiency of the missing data treatment methods depends on the missing mechanism. Consequently, Little and Rubin [14] have classified missing data into three categories:

Missing Completely at Random (MCAR): When given the variables X and Y, the probability of response dependents on X but not on Y.

Missing at Random (MAR): The probability of response independence exists between X and Y. MCAR data exhibits a high level if randomness than does MAR.

Non-ignorable: The probability of response depends on variables X and possibly on variable Y.

In practice it is usually difficult to meet the MCAR assumption. Most missing data methods are applied upon the assumption of MAR. And in correspondence to Kim [13], "Non-ignorable missing data is the hardest condition to deal with, but unfortunately, the most likely to occur as well". In

this paper, our experiments are based on the missing mechanisms MAR and MCAR.

Currently, machine learning based methods for imputing missing values include auto associative neural network [21], decision tree imputation [24], case-wise deletion [15], lazy decision tree [6], dynamic path generation [29]. But, these methods are not completely satisfactory ways to handle missing value problems because these methods perhaps destroy the original distribution of dataset during the process of imputing. Moreover, some methods in machine learning (such as C4.5) usually only handle with the discrete value. In these methods, continuous attributes are discretized before being processed, which may lose the true characteristic during the converting process from the continuous values to discretize ones, and the imputation result of those methods may destroy the original distribution of dataset.

From the data structure, commonly used imputation methods for missing values can be classified into parametric and non-parametric regressions. The parametric regression imputation is superior if a dataset can be adequately modeled parametrically, or if users can correctly specify the parametric forms for the dataset. If the model is misspecified (in fact, in real application, it is usually impossible for users to know the distribution of the real dataset), the estimation of parametric method may be highly biased, and then optimal control factor settings may be miscalculated.

Non-parametric imputation offer a nice alternative if users have no idea on the actual distribution of a dataset. Non-parametric imputation can provide superior fits by capturing structure in the dataset (note that a misspecified parametric model cannot), which is originally developed for situations with large sample sizes. In practice, however, non-parametric imputations often suffer from the curse of dimensionality in high dimensions, and in small sample settings, non-parametric fitting techniques may fit irregularities if the data are too closely [10].

While much work focuses on modeling data by parametric and nonparametric approaches, Engle et al. [5] have studied the semi-parametric model. In this case, data are based on monthly electricity sales y_i for four cities, the monthly price of electricity x_1, income x_2, and average daily temperature t. They modeled the electricity demand y as the sum of a smooth function g of monthly temperature t, and a linear function of x_1 and x_2, as well as 11 monthly dummy variables $x_3, \ldots, x_{13}$. Their model is

$$y = \sum_{j=1}^{13} \beta_j x_j + g(t) = X^T \beta + g(t)$$

Pin and James [20] have designed a semi-parametric conditional median as a robust alternative to the parametric conditional mean to estimate the gasoline demand function. Their approach protects against data and specification errors and may yield a more reliable basis for public policy decisions that depend on accurate estimates of gasoline demand.

Recently, Millimet et al. [17] has shown with data for US states that parametric modeling can be rejected in favor of a semi-parametric estimator, which does not impose any a priori restriction on the functional form of the relationship. Pickle et al. [19] compared the parametric and nonparametric methods, and present a semi-parametric for modeling which combine parametric and nonparametric function to improve the quality of both the mean and variance models. The resulting semi-parametric estimates have smaller bias and variance and result in a better understanding of the process at hand.

The above methods are based on an assumption that data are complete. Wang et al. [28] have developed inference tools in a semi-parametric partially linear regression model with missing response data. They have used a deterministic semi-parametric regression imputation with a view to avoid the curse of dimensionality. Based on the complete data after imputation, they make inference only for the mean of the response variable Y. Using a deterministic semi-parametric regression imputation method, while missing values in a dataset are replaced with only the mean of all the corresponding known values in the dataset, Wang and Rao [26] have showed that the deterministic imputation method performances well in making inference for the mean of Y.

Different from the above work, at first, we will propose an efficient random/stochastic semi-parametric regression imputation under the missing mechanisms MCAR and MAR. Then we will make an optimal inference for the response variable Y on RMSE, distribution function, and q-th quantile of Y. Using a stochastic semi-parametric regression imputation method, each of missing values in a dataset is replaced with the mean of all the corresponding known values in the dataset, plus a random value. Qin and Rao [22] have showed that one must use random imputation method in making inference for distribution functions and quantiles of Y. Our experimental results will demonstrate that stochastic semi-parametric regression imputation methods are much better than deterministic semi-parametric regression imputation methods on RMSE, distribution and quantiles.

3 Stochastic semi-parametric regression imputation method

In this section we present our data preprocessing in Section 3.1. And then we construct a kernel-based semi-parametric imputation for missing data in Section 3.2. In Section 3.3 we talk about the imputed values of Y, and the choice of bandwidth of kernel method and algorithm analysis are presented in Section 3.4.

3.1 Data preprocessing

A central concern is that the unit of an attribute can be very different from that of another attribute. For example, in a relation database, income of the inhabitants can take up values anywhere between 500 and 250,000, whereas the ratio of the employment content ranges from 0 to a maximum of 100%. Generally, the result is usually prone to the data with bigger magnitude, i.e., a unit difference in the ratio of the employment is expected to be more significant than the same unit difference in income of the inhabitants. To avoid this bias, the data in a database is transformed and normalized before data clustering and missing value imputation based on kernel functions in our paper.

Normalization is particularly useful when using kernel functions because normalization helps prevent attributes with initially largely ranges from outweighing attributes with initially smaller ranges. There are many methods for data normalization, for example, Min-Max normalization, z-score normalization and normalization by decimal scaling [8].

In this paper we first transform all input attributes to obtain temporary variables with distribution having zero mean and standard deviation of 1 using the following formula:

$$a_{ij(\text{temp})} = [(a_{ij}) - \bar{a}_j]/\sigma(a_j) \tag{3.1.1}$$

where a_{ij} represents the value of the jth attribute of the ith instance, $\bar{a}_j$ and $\sigma(a_j)$ represent the mean and standard deviation of the observed values of the jth attribute in the reference data set. And

$$a_{ij(\text{trans})} = a_{ij(\text{temp})}\{\text{MAX}[\text{range}(a_{j=1(\text{temp})}), \ldots, \text{range}(a_{j=x(\text{temp})})]\}/\text{range}(a_{j(\text{temp})}) \tag{3.1.2}$$

where $a_{j(\text{temp})}$ represents the data of the jth attributes normalized using formula (3.1.1); and $a_{ij(\text{trans})}$ represents the final transformed value of the jth attribute of the ith instance that is to be used as an input.

3.2 Constructing a semi-parametric regression model

A general semi-parametric regression model is as follows.

$$Y_i = X_i^T \beta + g(T_i) + \varepsilon_i \tag{3.2.1}$$

where the Y_i's are i.i.d (independent identically distributed) scalar response variables, the X_i's are i.i.d d-dimensional random covariate vectors, the T_i's are i.i.d d^*-dimensional random covariate vectors, the function $g(\cdot)$ is unknown, and the model error ε_i are i.i.d random errors with mean 0 and unknown finite variance σ^2 (in our paper, we regard the unknown finite variance as 1).

We consider the case where some Y values in a sample size n may be missing, but X and T are observed completely. That is, we obtain the following incompletely observations:

$$(Y_i, \delta_i, X_i, T_i), \quad i = 1, 2, \ldots, n$$

from model (3.2.1). Where all the X_i's and T_i's are observed and $\delta_i = 0$ if Y_i is missing and $\delta_i = 1$ otherwise.

We assume that Y is missing at random (MAR). The MAR assumption implies that δ and Y are conditionally independent given X and T. That is,

$$P(\delta = 1|Y, X, T) = P(\delta = 1|X, T)$$

In practice, the MAR assumption is usually justified in the nature of experiments, especially when it is speculated that missing Y mainly depends on X. The MAR has received most attentions theoretically and practically as it describes the natural practical case. MCAR is a stronger assumption than MAR (i.e. MCAR is a special case of MAR). MCAR implies that the probability of missing a value is the same for all variables X and T.

Let $r = \sum_{i=1}^{n} \delta_i, m = n - r$ Denote the set of respondents and non-respondents as S_r (all data are observed in this sets) and S_m (there is missing in Y, but all data in X are observed in S_m), respectively. Let K be a symmetric probability density function and let $h = h_n$ be a bandwidth sequence that decreases toward 0 as the sample size n increases toward $+\infty$. From (3.2.1), we have:

$$Y_i - X_i^T \beta = g(T_i) + \varepsilon_i. \quad i = 1, \ldots, r \tag{3.2.2}$$

Assuming β is known, we have a kernel estimator $\hat{g}(t)$ for $g(t)$ based on the completely observed data:

$$\hat{g}(T_i) = \frac{\sum_{j=1}^{n} \delta_j K\left(\frac{(T_i - T_j)}{h}\right)(Y_j - X_j\beta)}{\sum_{j=1}^{n} \delta_j K\left(\frac{(T_i - T_j)}{h}\right) + n^{-2}}. \quad i = 1, \ldots, r \tag{3.2.3}$$

where the term n^{-2} is introduced to avoid the case that the denominator vanishes; $K(\cdot)$ is called kernel function. There are some widely used kernel functions in semi-parametric inference, i.e. the Gaussian kernel (standard normal density function)

$$K(\cdot) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right), \quad t \sim N(1, 1)$$

and a polynomial kernel

$$K(\cdot) = \begin{cases} \frac{15}{16}(1 - t^2 + t^4), & |t| \le 1 \\ 0, & \text{otherwise} \end{cases}$$

In practice, there is not any significant difference using these kernel functions under the MAR and MCAR assumptions. In this paper, we use the polynomial kernel in our experiments. We will discuss the choosing of bandwidth in kernel method later in Section 3.4.

Using $\hat{g}(T_i)$ to replace $g(T_i)$ in (3.2.2), we obtain:

$$Y_i - X_i^T \beta \approx \frac{\sum_{j=1}^n \delta_j K\left(\frac{(T_i - T_j)}{h}\right)(Y_j - X_j\beta)}{\sum_{j=1}^n \delta_j K\left(\frac{(T_i - T_j)}{h}\right) + n^{-2}}, \quad i \in s_r. \tag{3.2.4}$$

Converting (3.2.4), we have

$$Z_i \approx U_i^T \beta, \quad i \in s_r \tag{3.2.5}$$

Where

$$Z_i = Y_i - \frac{\sum_{j=1}^n \delta_j Y_j K\left(\frac{(T_i - T_j)}{h}\right)}{\sum_{j=1}^n \delta_j K\left(\frac{(T_i - T_j)}{h}\right) + n^{-2}},$$

$$U_i = X_i - \frac{\sum_{j=1}^n \delta_j X_j K\left(\frac{(T_i - T_j)}{h}\right)}{\sum_{j=1}^n \delta_j K\left(\frac{(T_i - T_j)}{h}\right) + n^{-2}}, \qquad i \in s_r. \tag{3.2.6}$$

According to the theory of linear regression model, β is estimated by (3.2.7):

$$\hat{\beta}_n = \left(\sum_{i=1}^n \delta_i U_i U_i^T\right)^{-1} \left(\sum_{i=1}^n \delta_i U_i Z_i\right). \tag{3.2.7}$$

where n is the sample size.

Combining with (3.2.3), the final estimator for $g(T_i)$ is given by

$$\hat{g}_n(T_i) = \frac{\sum_{j=1}^n \delta_j K\left(\frac{(T_i - T_j)}{h}\right)(Y_j - X_j\hat{\beta}_n)}{\sum_{j=1}^n \delta_j K\left(\frac{(T_i - T_j)}{h}\right) + n^{-2}}. \tag{3.2.8}$$

3.3 Imputation missing values in Y

Let $Y_i^{(D)}$ and $Y_i^{(R)}$, $i \in s_m$ be the imputed values for the missing data based on deterministic and random semi-parametric imputation methods, respectively. Deterministic semi-parametric imputation [27] uses $\hat{g}_n(T_i)$ as the imputed value, i.e.

$$Y_i^{(D)} = X_i^T \hat{\beta}_n + \hat{g}_n(T_i), \quad i \in s_m$$

In this paper, we construct the random semi-parametric regression imputation and regard $Y_i^{(R)} = X_i^T \hat{\beta}_n + \hat{g}_n(T_i) + \varepsilon_i^* = Y_i^{(D)} + \varepsilon_i^*$ $(i \in s_m)$ as the imputed values, which have same convergence as the deterministic method, where $\{\varepsilon_i^*\}$ is randomly obtained from $\{Y_i - X_i^T \hat{\beta}_n - \hat{g}_n(T_i), \quad i \in s_r\}$.

Denote $Y_{D,i} = \delta_i Y_i + (1 - \delta_i) Y_i^{(D)}$, $Y_{R,i} = \delta_i Y_i + (1 - \delta_i) Y_i^{(R)}$, $i = 1, \ldots, n$
which are 'complete' data based on the above imputations.

Then, we make inference for the response variable Y such as RMSE (Root Mean Square Error), distribution function $(\theta = F(y))$ and quantile (θ_q) after missing-data are imputed to present our performance of our algorithm. Based on the complete data after imputation, the standard estimators for the parameters can be constructed as follows.

The standard estimators of $\theta = F(y)$ under random and deterministic imputation are given respectively by

$$F_R(y) = \frac{1}{n}\sum_{i=1}^n I(Y_{R,i} \le y)$$

$$F_D(y) = \frac{1}{n}\sum_{i=1}^n I(Y_{D,i} \le y)$$

The standard estimators of $\theta_q = F^{-1}(q)$ under random and deterministic imputation are given respectively by

$$\hat{\theta}_q^{(R)} = \inf_u\{F_R(u) \ge q\} = F_R^{-1}(q)$$

$$\hat{\theta}_q^{(D)} = \inf_u\{F_D(u) \ge q\} = F_D^{-1}(q)$$

As well, we design the other experiment to evaluate the performance of our algorithm, i.e. the accuracy of prediction was measured using the Root Mean Square Error (RMSE) to present the performance between our presented imputation method and the existed methods, the RMSE is as follows:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^m (e_i - \tilde{e}_i)^2}. \tag{3.3.1}$$

where e_i is the original value; $\tilde{e}_i$ is the estimated value, and m is the total number of predictions. The RMSE is more, the least the prediction accuracy is.

3.4 Bandwidth choosing and algorithm analysis

Kernel method can be decomposed into two parts: one for the calculation of the kernel and another for bandwidth choice.

Silverman [25] stated that one important factor in reducing the computer time is the choice of a kernel that can be calculated very quickly. Having chosen a kernel that is efficient to compute, one must then choose the bandwidth. Silverman [25] pointed out that the choice of bandwidth is much more important than the choice of kernel function.

Generally, a small value of bandwidth h can make the estimate look 'wiggly' and show spurious features, whereas a big value of h will lead to an estimate that is too smooth, in the sense, that it is too biased and may not reveal structural features. There is no generally accepted method for choosing the bandwidths. Methods currently available include 'subjective choice' and automatic methods such as the "plug-in", 'cross-validation' (CV), and 'penalizing function' approaches. In this paper, we use the method of cross-validation to minimize the approximate mean integrated square error (AMISE) of $\hat{g}(T_i)$ for a given sample of data.

Define the CV function as

$$CV = \sum_{i=1}^{n} (Y_i - X_i\hat{\beta}_n - \hat{g}_{-i}(T_i))^2$$

where $\hat{g}_{-i}(T_i)$ denotes the 'leave-one-out' kernel estimator of $g(T_i)$, i. e.

$$\hat{g}_{-i}(T_i) = \frac{\sum_{j\neq i} K\left(\frac{(T_i - T_j)}{h}\right)(Y_j - X_j\hat{\beta}_n)}{\sum_{j\neq i} K\left(\frac{(T_i - T_j)}{h}\right) + n^{-2}}$$

While the complexity of the kernel method is $O(mn^2)$, where n is the number of instances of the dataset, m is the number of attributes, so the algorithm complexity of our method is $O(kmn^2)$ (k is the number of missing values).

4 Experimental results

Our methodology consists of two phases: (1) filling up missing values in a dataset based on stochastic semi-parametric regression imputation method in Section 3; (2) evaluating the quality of the imputed datasets, where we compare the performance of our stochastic semi-parametric regression imputation with the deterministic method in terms of the imputation in Section 4.1, and present the performance of our algorithm in real dataset in Section 4.2. We conduct our experiments using a DELL Workstation PWS650 with 2G main memory, 2.6G CPU, and WINDOWS 2000.

4.1 Simulation model

We conducted a series of simulation studies on the finite sample performance of the deterministic and random imputations in distribution function $\theta = F(y)$ for fixed y, quantile $\theta_q = F^{-1}(q)$ and RMSE. The performance is measured in terms of the mean squared errors (MSE) of estimators, i.e. the average squared errors over repeated time of simulations. For this purpose, we took K as the polynomial density function.

According to (3.2.1), we used model:

$$\beta = 1.5,$$

and

$$g(t) = 3.2t^2 - 1 \quad \text{if } t \in [0, 1], \quad g(t) = 0, \quad \text{otherwise.}$$

We generated $X_i s$ from the normal distribution $N(1, 1)$ and $\varepsilon_i s$ from the standard normal distribution $N(0, 1)$, and the following two cases of response probabilities under the MAR and MCAR assumptions from [26]:

Case 1 (MAR):
$P_1(x, t) = P(\delta = 1|X = x) = 0.8 + 0.2(|x - 1 + |t - 0.5|)$, if $|x - 1| + |t - 0.5| \leq 1$, and $=0.95$, elsewhere.

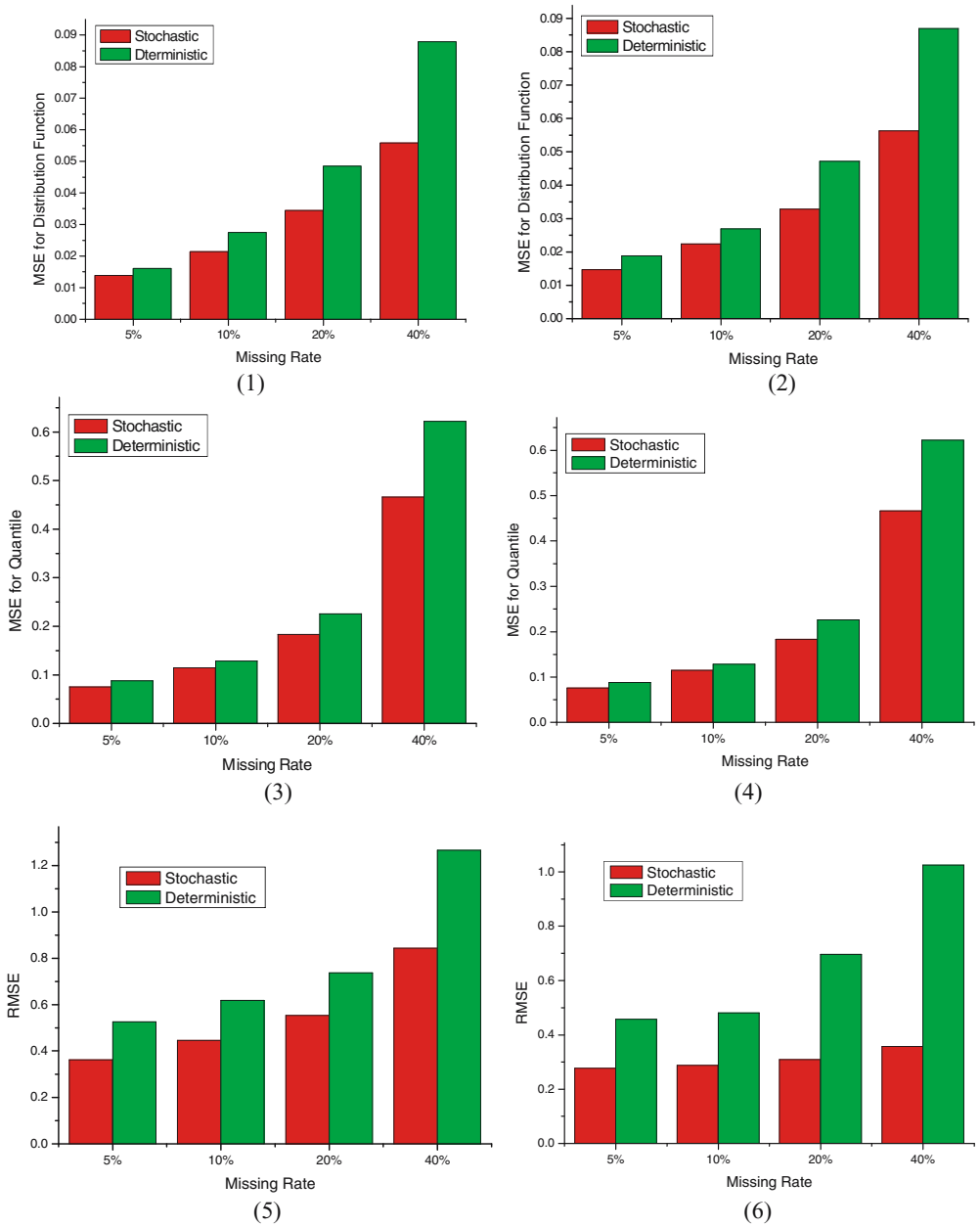
Case 2 (MCAR):
$P(\delta = 1|X = x, T = t) = 0.6$, for all x and t.

For each of the two cases, we generated 1,000 (repeated time) random samples of incomplete data $\{X_i, Y_i, \delta_i, i = 1, \ldots, n\}$ for $n = 100$ from the models and specified response probability function.

Figure 1 presents the performance of two imputation methods for MSE of distribution function with missing rate 5%, 10%, 20% and 40% under MCAR and Fig. 2 presents the results under MAR. Figure 3 presents the performance of two imputation methods for quantile with missing rate 5%, 10%, 20% and 40% under MCAR and Fig. 4 presents the results under MAR. Stochastic refers to the stochastic semi-parametric regression imputation method and deterministic denotes the deterministic semi-parametric regression imputation method from Figs. 1–4. Figure 5 presents the result of RMSE for two imputation methods with missing rate 5%, 10%, 20%, and 40% under MCAR and Fig. 6 presents the results under MAR

Figures 1–6 reveal the following facts:

1. About the performance of the distribution function, stochastic semi-parametric regression imputation is uniformly better than the deterministic semi-parametric regression imputation for various missing rate as shown in Figs. 1 and 2 under the missing mechanism of MACR or MAR; Stochastic imputation is almost uniformly better than the deterministic imputation in making inference



on the quantile of Y for different response rates as shown in Figs. 3 and 4. Stochastic semi-parametric imputation method is also significantly better than the deterministic method about the accuracy of prediction from the performance of RMSE.

2. Comparing to missing rate, we can see that the performance is better when the response rate is higher about the distribution function, quantile and RMSE.

4.2 Real dataset

We considered the real data set given in [9, 18]. The data give the normal average January minimum temperature in degrees Fahrenheit (Denoted as *JanTemp*) with the latitude (*Lat*) and longitude (*Long*) of 56 U.S. cities. For each year from 1931 to 1960, the daily minimum temperatures in January were added together and divided by 31. Then, the averages for each year were averaged over the 30 years. The data set is also available on:

http://lib.stat.cmu.edu/DASL/Datafiles/USTemperatures.html.

We suppose the dependent variable (Y) is *JanTemp* and the independent variable (X) is *Lat*. Our experiment present that the value of significant probability of the correlation between

the *JanTemp* and *Lat* is 0 in software SPSS, after removing the effects of *Lat*, we get the value of significant probability of the correlation between the *JanTemp* and *Long* is 0.861, these result show there is an evidently linear relationship between *JanTemp* and *Lat*, the linear relationship between the *JanTemp* and *Long* is not clearly. To apply our method to these real data, we denote the variables for *JanTemp*, *Lat* and *Long* to be Y, X and T respectively. We suppose that Y, X and T satisfy the semi-parametric model (3.2.1).

Note that the original data set given by Peixoto [18] is complete. Inference on the distribution function of Y or quantile of Y with the complete data set doesn't depend on the model assumption and covariables X and T (as the data for Y is complete and the standard statistical procedures can by applied directly to make inference for the parameters of Y). This just provides us a standard to compare our methods with other methods to handle missing data. In this section, we also compare our stochastic semi-parametric regression imputation with the deterministic semi-parametric regression imputation, the non-parametric kernel regression imputation methods and linear regression methods.

We used all the 56 data and random deleted 6, 14 or 23 Y values (Missing Rate is almost 10%, 20% or 40% respectively) and the repeated times are 1000. The deletion mechanisms are designed to be MAR and MCAR same as the Section 4.1. We make inference on the distribution function $\theta = F(y)$ for fixed y, quantile $\theta_q = F^{-1}(q)$ and RMSE comparing our stochastic semi-parametric regression imputation estimator with deterministic semi-parametric regression imputation, non-parametric model and linear model.

When making inference based on nonparametric kernel regression imputation estimator, the kernel function $K(t)$ and the deletion mechanism were taken to the same as in Section 4.1. For calculation of $\hat{g}_n(X, T)$ (which is the estimator of $g(X, T)$ in the nonparametric regression model $Y = g(X, T) + \varepsilon$) based on [34], it was taken to be

$$\hat{g}_n(X,T) = \frac{\sum_{i=1}^{n} \delta_i Y_i K_1\left(\frac{X-X_i)}{h}\right) K_2\left(\frac{T-T_i}{h}\right)}{\sum_{i=1}^{n} \delta_i K_1\left(\frac{X-X_i)}{h}\right) K_2\left(\frac{T-T_i}{h}\right) + n^{-2}}$$

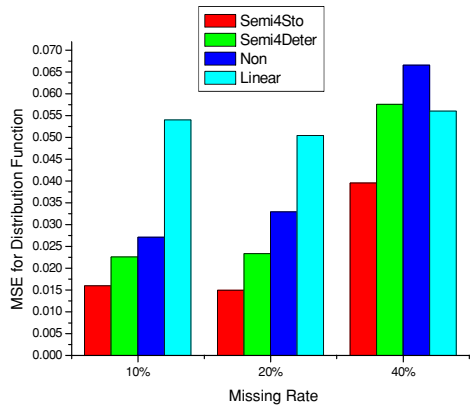
where $K_1(\cdot)$, $K_2(\cdot)$ are the same as K in the simulation study, here the selection of bandwidth h is same as that in Section 3.4.

Due to the evidently linear relationship between *JanTemp* and *Lat*, we assume the linear multiple-regression among *JanTemp*, *Lat* and *Long*, and then we construct an experiment about multiple linear regression imputation comparing with the non-parametric and semi-parametric model.

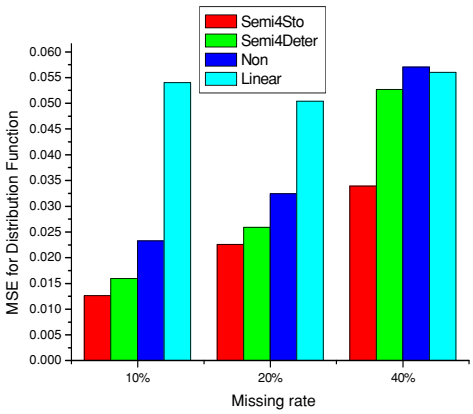
Figure 7 present the performance of four imputation methods for distribution function with missing rates 10%, 20% and 40% under MCAR, and Fig. 8 is under the missing mechanism of MAR. Figures 9 present the performance of four imputation methods for quantile with missing rates 10%, 20% and 40% under MCAR, and Fig. 10 is under the missing mechanism of MAR. Figure 11 present the performance of four imputation methods for RMSE with missing rates 10%, 20% and 40% under MCAR, and Fig. 12 is under the missing mechanism of MAR. The '*Semi4Sto*', '*Semi4Deter*', '*Non*', '*Linear*' refer to as the imputation method stochastic semi-parametric regression imputation method, deterministic semi-parametric regression imputation method, non-parametric regression imputation method and linear regression method respectively.

Figures 7–12 reveal the following facts:

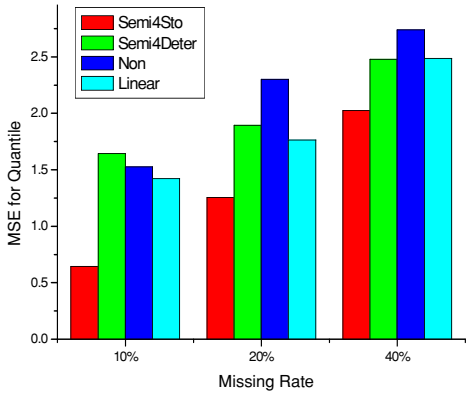
1. From Figs. 7–12, we can see that the performances of two imputation methods based on semi-parametric models are similar with the simulations results shown before. The stochastic imputation method is basically better than the deterministic imputation method in making inference on the distribution function, quantile and RMSE of the response variable.
2. The performances based on the two semi-parametric models are significantly better than the nonparametric model and linear model as there is some linear relation between the covariates and the response variable because the semi-parametric model is capable to take this information into account.
3. Comparing to missing rate, we can see that the performance of our stochastic semi-parametric regression imputation method is better when the response rate is higher about the distribution function, quantile and RMSE under the missing mechanism of MCAR or MAR, and the result show the performance of the three methods besides our stochastic semi-parametric imputation method fluctuate when the missing rate is increase from Figs. 7–12, such as, the linear model is worst.
4. We get the result from the performance that the best efficient method is our stochastic semi-parametric regression imputation method, then the better is deterministic and non-parametric, and the worst is linear method based on the real data. We also get a conclusion: we had better use semi-parametric regression imputation method to patch up the missing value when we have a little information about the missing attribute variables and the observed attributes variable, such as we know the linear relationship between the dependent variable and one of the independent variables.



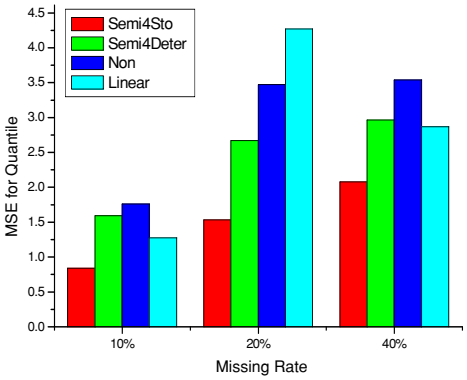
(7)



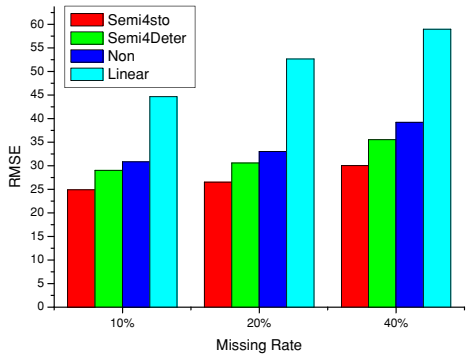
(8)



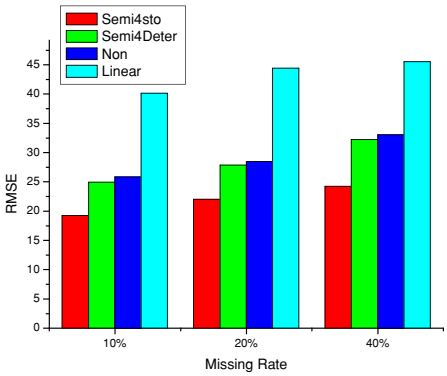
(9)



(10)



(11)



(12)

5 Conclusions and future work

In many practical situations, either parametric models, or non-parametric models are not capable enough to capture the underlying relation between the response variable and its associated covariates when we have a little priori knowledge about the real dataset. In this case, we have argued to use a semi-parametric model when having priori knowledge during handling with missing values. In this semi-parametric regression setting, we have shown that the stochastic/random regression imputation works well in making inference on all the response variables. It has also illustrated that the deterministic regression imputation is not well for the distribution function, quantiles of Y and RMSE than the stochastic method. That is, when we need to make inference on the distribution function, quantiles of Y or RMSE, we recommend users to use random imputation.

References

1. Allison P (2001) Missing data. Sage Publication, Inc
2. Cios K, Kurgan L (2002) Trends in data mining and knowledge discovery. In: Pal N, Jain L, Teoderesku N (eds) Knowledge discovery in advanced information systems. Springer

3. Clifton C (2003) Change detection in overhead imagery using neural networks. Appl Intell 18(2):215–234
4. Dempster et al (1983) Incomplete data in sample surveys. In: Madow WG, Olkin I, Rubin D (eds) Sample surveys Vol.: Theory and annotated bibliography, New York, NY, Academic Press, pp 3–10
5. Engle RF et al (1986) Semiparametric estimates of the relation between weather and electricity sales. J Am Statist Assoc 81(394), Applications.
6. Friedman JH, Khavi R, Yun Y (1996) Lazy decision trees. In: Proceedings of the 13th national conference on artificial intelligence, AAAI Pres/MIT Press, pp 717–724
7. Ghahramani et al (1997) Mixture models for Learning from incomplete data. In: Greiner R, Petsche T, Hanson SJ (eds) Computational learning theory and natural learning systems, Volume IV: Making learning systems practical, Cambridge, MA, The MIT Press, pp 67–85
8. Han J, Kamber M (2000) Data mining concepts and techniques. Morgan Kaufmann Publishers
9. Hand D et al (1994) A handbook of small data sets. London, Chapman & Hall, pp 208–211
10. Hoti F, Holmstrom L (2004) A semiparametric density estimation approach to pattern classification. Patt Recog 37:409–419
11. Hu X (2005) A data mining approach for retailing bank customer attrition analysis. Appl Intell 22(1):47–60
12. Kaya M, Alhajj R (2006) Utilizing genetic algorithms to optimize membership functions for fuzzy weighted association rule mining. Appl Intell 24(1):7–15
13. Kim Y (2001) The curse of the missing data. In: http://209.68.240.11:8080
14. Little R, Rubin D (2002) Statistical analysis with missing data (2nd edn.). John Wiley and Sons, New York
15. Liu WZ, White AP, Thompson SG, Bramer MA (1997) Techniques for dealing with missing values in classification. In: IDAL97, vol 1280 of Lecture notes, pp 527–536
16. Ramoni M (1997) Learning Bayesian networks from incomplete databases. Technical report kmi-97-6, Knowledge Media Institute, The Open University
17. Millimet D, List J, Stengos T (2003) The environmental kuznets curve: Real progress or misspecified models? Rev Econ Stat 85(4):1038–1047
18. Peixoto J (1990) A property of well-formulated polynomial regression models. Am Stat 44:26–30
19. Pickle S et al (2005). Robust parameter design: a semi-parametric approach. In: http://www.stat.vt.edu/tech_reports/VTTechReport05-7.pdf
20. Pin T, James L (1999) The elasticity of demand for gasoline: a semi-parametric analysis. In http://uiuc.edu/∼ng/working/gas.ps
21. Pyle D (1994) Data preparation for data mining. Morgan Kaufmann Publishers, Inc
22. Qin YS, Rao JNK (2004) Confidence intervals for parameters of the response variable in a linear model with missing data. Technique Report
23. Quinlan JR (1989) Unknown attribute values in induction. In: proc. 6th int' workshop on machine learning, Ithaca, pp 164–168
24. Quinlan JR (1993) C4.5: Programs for machine learning. Morgan Kaufmann, San Mateo, USA
25. Silverman B (1986) Density estimation for statistics and data analysis. Chapman and Hall, New York
26. Wang Q, Rao JNK (2002a) Empirical likelihood-based inference in linear models with missing data. Scand J Statist 29:563–576
27. Wang Q, Rao J (2002b) Empirical likelihood-based inference under imputation with missing response. Ann Statistics 30:563–576
28. Wang Q, Hardle W (2004) Semiparametric regression analysis with missing response at random. J Am Statistical Assoc 99
29. White AP (1987) Probabilistic induction by dynamic path generation in virtual trees. In: Bramer MA (ed) Research and development in expert systems III. Cambridge, Cambridge University Press, pp 35–46
30. Zhang C, Yang Q, Liu B (2005) Intelligent data preparation. IEEE Trans Knowl Data Eng 17(9):1163–1165
31. Zhang C, Zhang S, Webb G (2003) Identifying approximate itemsets of interest in large databases. Appl Intell 18:91–104
32. Zhang S, Zhang C, Yang Q (2004) Information enhancement for data mining. IEEE Intell Syst 19(2):12–13
33. Zhang S, Qin ZX, Ling CX, Sheng SL (2005) Missing is useful: missing values in cost-sensitive decision trees. IEEE Trans Knowl Data Eng 17(12):1689–1693
34. Zhang S et al (2006) Optimized parameters for missing data imputation. In: Proceedings of PRICAI 2006, Guilin, China, August 7–11, 2006 Proceedings, pp 1010–1016