Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators

Pier Luigi Conti\textsuperscript{a}, Daniela Marella\textsuperscript{a,}\textsuperscript{*}, Mauro Scanu\textsuperscript{b}

\textsuperscript{a} Università di Roma “La Sapienza”, Italy
\textsuperscript{b} Istituto Nazionale di Statistica, Italy

\textbf{A B S T R A C T}

A new matching procedure based on imputing missing data by means of a local linear estimator of the underlying population regression function (that is assumed not necessarily linear) is introduced. Such a procedure is compared to other traditional approaches, more precisely hot deck methods as well as methods based on kNN estimators. The relationship between the variables of interest is assumed not necessarily linear. Performance is measured by the matching noise given by the discrepancy between the distribution generating genuine data and the distribution generating imputed values.

\section{1. Introduction}

In several contexts, e.g. official statistics (D'Orazio et al., 2002, 2006), marketing (Räessler, 2002), genetics (as for the data sets in repositories like genenetwork.org), data files coming from different sources are frequently available at a moderate cost. Each data file contains the values of some of the variables of interest. This is a serious limitation, when one is interested in the joint analysis of variables that are not jointly observed.

The statistical matching problem consists in constructing a complete synthetic data file, where all the variables of interest are present. In a sense, this is a purely “descriptive” objective, representing the multivariate joint distribution, with the aim to create a data set available to end-users.

The synthetic data set is constructed by using imputation techniques. As a consequence the joint distribution of the variables of interest in the synthetic data file does not generally coincide with the genuine distribution. This discrepancy is the matching noise. From an end-user perspective, the smaller the matching noise, the better the reconstructed data file.

Different techniques have been proposed in the literature for tackling the statistical matching problem, among them an important role is played by hot deck methods, as well as kNN methods. Their properties are studied in Paass (1985) and Marella et al. (2008), where both theoretical and simulation results are obtained. In this paper we go further by introducing new nonparametric matching techniques based on local linear regression, that are compared to existing ones.

The paper is organized as follows. In Section 2 the main technical aspects are briefly introduced. In Section 3 a class of nonparametric imputation procedures are described, including the method based on the local linear estimator. In Section 4 the matching noise (for imputation based on local linear regression estimators) is formally evaluated. Finally, in Section 5 a simulation study is implemented.

\textsuperscript{*} Corresponding address: Università di Roma “La Sapienza”, piazzale Aldo Moro 5, 00185 Roma, Italy. Tel.: +39 06 49910657; fax: +39 06 4959241.
E-mail address: daniela.marella@uniroma1.it (D. Marella).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.

2. Notation and technicalities

With no loss of generality, in order to accomplish the goal described in the introduction, we consider here a simplified, technically affordable case, where only two variables are involved. However the main ideas of Section 3.2.2, as well as results in Propositions 1 and 2, can be easily extended to more general cases.

Let \((X, Z)\) be a bivariate random variable (r.v.) with density function \(f(x, z)\), and let \(A, B\) be two independent samples of \(n_A\) and \(n_B\) i.i.d. records from \((X, Z)\), respectively, where \(n_A\) and \(n_B\) are fixed in advance by design. The first \(n_A\) records have \(Z\) missing while the last \(n_B\) records are complete. Hence,

\[
\begin{align*}
\{x^A\} &= \{x_1^A, \ldots, x_{n_A}^A\}, \\
\{x^B, z^B\} &= \{(x_1^B, z_1^B), \ldots, (x_{n_B}^B, z_{n_B}^B)\},
\end{align*}
\]

are the observed values in \(A\) and \(B\), respectively. This is the typical situation in statistical matching where missingness is induced by survey design and can be considered deterministic.

In this paper we mainly focus on a particular imputation procedure of \(Z\) in \(A\), based on the nonparametric estimation of the regression function via local linear estimators. For this imputation procedure, the matching noise is studied by both theoretical and simulation approaches. Furthermore, the proposed procedure is compared with other nonparametric imputation procedures. The most popular ones are those based on hot deck, both theoretical and simulation approaches. Furthermore, the proposed procedure is compared with other nonparametric estimators. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.07.041

3. Nonparametric imputation procedures

In order to appropriately impute missing data, the model that generates imputations should equal the data generating model: the distribution of \((X, Z)\) should coincide with the distribution of \((X, Z)\). Either implicitly or explicitly, the model that generates imputations is estimated from the observed data. In the case of the data sets (1), the joint \((X, Z)\) distribution is preserved when \(Z\) is imputed according to the genuine conditional distribution of \(Z\) given \(X\). This conditional distribution must be estimated only on the basis of sample \(B\) (see Rubin (1974)).

In the sequel (Sections 3.1, 3.2 and 3.2.1) a short description of widely used nonparametric imputation techniques is given. Section 3.2.2 illustrates another imputation technique based on the local polynomial estimation of the regression function. Most of these techniques are based on the concept of neighbour. Formally, for each \(a = 1, \ldots, n_A\), let \(b(a) = (b_1(a), \ldots, b_k(a))\) be the labels of the \(k \geq 1\) nearest neighbours of \(x_a\) in \(B\), such that

\[
d(x_a^A, x_{b_j(a)}^B) \leq d(x_a^A, x_{b_{j+1}(a)}^B), \quad j = 1, \ldots, k - 1,
\]

and

\[
d(x_a^A, x_{b_k(a)}^B) \leq d(x_a^A, x_{b_1(a)}^B), \quad \forall b \notin \{b_1(a), \ldots, b_k(a)\},
\]

where \(d(\ldots)\) is the Euclidean distance. Let \(x_{b(a)}^B = (x_{b_1(a)}^B, x_{b_2(a)}^B, \ldots, x_{b_k(a)}^B)\) and \(z_{b(a)}^B = (z_{b_1(a)}^B, z_{b_2(a)}^B, \ldots, z_{b_k(a)}^B)\) be the vectors of corresponding \(X\) and \(Z\) values, respectively.

3.1. kNN random hot deck and distance hot deck

Once the \(k\) nearest neighbours of \(x_a^A, x_{b(a)}^B\), are obtained, one could impute the missing \(z_a^A\) by randomly choosing a label \(\tilde{b}(a)\) among \(b_j(a), j = 1, \ldots, k\), and in taking imputed values

\[
z_a^A = z_{\tilde{b}(a)}^B, \quad a = 1, \ldots, n_A.
\]
A generalized version of this approach is in Aluja-Banet et al. (2007). A value is taken at random assuming different probabilities of selection for the donor records: observations close to \( x_a \) have higher probabilities than those further away.

When \( k = 1 \), this imputation method reduces to distance hot deck. Imputed data are obtained as:

\[
\tilde{z}_a^k = z_{b_j(a)}, \quad a = 1, \ldots, n_A.
\]

(3)

In other words, each record in \( A \) is matched with the closest record in \( B \).

Note that distance hot deck can be defined in two versions. The one in Eq. (3) is distance hot deck with replacement (i.e. each donor can be used more than once). Distance hot deck without replacement (each donor can be used at most once) is not considered here.

3.2. Methods based on nonparametric regression function

Since Yates (1933), when \( X \) and \( Z \) are continuous a very important role has been played by the regression function of \( Z \) on \( X \). More precisely, a linear regression function is assumed. A simple (and natural, as well) idea to impute missing data consists in using a nonparametric estimator of the (not necessarily linear) regression function of \( Z \) on \( X \) (see, for instance, Chu and Cheng (1995) and Nielsen (2001)). Suppose that \( X \) and \( Z \) are related through the relationship

\[
Z = m(X) + \epsilon,
\]

(4)

where \( m(x) = E[Z|X = x] \) is the regression function of \( Z \) given \( X \) and \( \epsilon = Z - m(X) \) is the error term, such that \( E[\epsilon|X = x] = 0 \) for every \( x \). For the sake of simplicity, in the sequel we will further assume that the errors are homoscedastic, i.e. \( \sigma^2 \) independent of \( x \). A simple idea to impute missing data \( z^A \) in the sample \( A \) could consist of the following steps.

1. Estimate the regression function \( m(x) \) by the sample \( B \). From now on, such an estimator will be denoted by \( \hat{m}^B(x) \).
2. Let

\[
\tilde{e}_b = \hat{m}^B(x_b^A) - \hat{m}^B(x_a^B), \quad b = 1, \ldots, n_B,
\]

(5)

be the corresponding residuals in \( B \).
3. Impute the missing \( z_a^A \)’s by

\[
\tilde{z}_a^A = \hat{m}^B(x_a^A) + \tilde{e}_a, \quad a = 1, \ldots, n_A,
\]

(6)

where \( \tilde{e}_a \) is drawn at random among \( \tilde{e}_1, \ldots, \tilde{e}_{n_B} \).

The rationale of step 1–3 is simple: at first model (4) is estimated by the complete sample \( B \), and then used to impute the missing data \( z_a^A \)’s in \( A \). According to Brick and Kalton (1996) this is a stochastic imputation method. Clearly, if estimated residuals \( \tilde{e}_b \) are omitted in (6), so that

\[
\tilde{z}_a^A = \hat{m}^B(x_a^A), \quad a = 1, \ldots, n_A,
\]

(7)

then the imputation method is deterministic. In the sequel, a short description of two imputation procedures based on estimating the regression function \( m(x) \) through the kNN estimator and the local linear regression estimator is given.

3.2.1. kNN methods

The kNN imputation method consists in estimating the nonparametric regression function \( m(x) \) by the kNN estimator. Formally, the regression function \( m(x) \) is estimated by the average of \( Z \) corresponding to the \( k \) nearest neighbours of \( x \). When \( x = x_a^A \):

\[
\hat{m}^B(x_a^A) = \frac{1}{k} \sum_{j=1}^{k} \tilde{z}_{b_j(a)}, \quad a = 1, \ldots, n_A.
\]

Deterministic imputation computed from the estimated nonparametric regression function is:

\[
\tilde{z}_a^A = \hat{m}^B(x_a^A), \quad a = 1, \ldots, n_A.
\]

(8)

The corresponding stochastic imputation is obtained by

\[
\tilde{z}_a^A = \hat{m}^B(x_a^A) + \tilde{e}_a, \quad a = 1, \ldots, n_A,
\]

(9)

where \( \tilde{e}_a \) is chosen at random from the residuals computed as in (5) on file \( B \). The key point in using the kNN estimator (8) is the choice of the parameter \( k \), that determines the amount of smoothing of \( z^B \)’s data. It plays a role similar to the bandwidth for kernel smoothers.

It can be shown (Paas, 1985; Cohen, 1991) that distance hot deck described in Section 3.1, Eq. (3), is equivalent to impute missing data through the kNN method, with \( k = 1 \). Such a procedure seems to be at first sight a deterministic technique, because residuals estimated as in Eq. (5) are null whenever \( x \) is equal to any of the \( n_B \) values \( x_b^A \) observed in \( B \). As a matter of fact this method imputes at the same time both the regression function and the residual. However, this does not mean that the matching noise is null. It can be proved (see Marella et al. (2008)) that the matching noise still affects this imputation approach for finite \( n_B \), although it becomes negligible for large \( n_B \).

3.2.2. Local polynomial estimator

As an alternative to kNN estimator, the local polynomial estimators (Fan and Gijbels, 1996) represent a simple and useful class of estimators of the regression function \( m(x) \). Suppose that \( m(x) \) possesses \( p + 1 \) derivatives, and denote by \( m^{(j)}(x) \) its \( j \)th derivative, \( j = 1, \ldots, p + 1 \). The basic idea consists in approximating \( m(t) \) locally by a polynomial of order \( p \):

\[
m(t) \approx m(x) + m^{(1)}(x)(t - x) + \cdots + \frac{1}{p!} m^{(p)}(x)(t - x)^p
\]

= \( \beta_0 + \beta_1(t - x) + \cdots + \beta_p(t - x)^p \).

Model (10) may be considered as a “usual” polynomial model on a local scale, with parameters \( \beta_0, \ldots, \beta_p \) depending on \( x \). They may be estimated by the weighted least squares method, which consists in minimizing the quantity:

\[
\sum_{b=1}^{n_\theta} \left( Z_b^B - \sum_{j=0}^{p} \beta_j(X_b^B - x)^j \right)^2 K_h \left( X_b^B - x \right),
\]

where \( K(\cdot) \) is a nonnegative weight function, \( K_h(t) = h^{-1} K(t/h) \), and \( h \) (the bandwidth) is a smoothing parameter determining the size of the neighbourhood of \( x \) used in estimating \( m(x) \).

Local polynomial estimators have been proved as particularly useful, and efficient as well. Their merits are thoroughly discussed in Fan and Gijbels (1996). In particular, when \( p = 0 \) the local polynomial estimator reduces to the Nadaraya–Watson estimator, that may be written as:

\[
\hat{m}_0^B(x) = \frac{\sum_{b=1}^{n_\theta} Z_b^B K_h \left( X_b^B - x \right)}{\sum_{b=1}^{n_\theta} K_h \left( X_b^B - x \right)}.
\]

When \( p = 1 \), the local polynomial estimator reduces to the local linear estimator, that may be written in the form:

\[
\hat{m}_1^B(x) = \frac{S_2^B(x) T_1^B(x) - S_1^B(x) T_2^B(x)}{S_2^B(x) S_0^B(x) - S_1^B(x)^2},
\]

where

\[
S_j^B(x) = \sum_{b=1}^{n_\theta} \left( x - X_b^B \right)^j K_h \left( X_b^B - x \right),
\]

\[
T_j^B(x) = \sum_{b=1}^{n_\theta} Z_b^B \left( x - X_b^B \right)^j K_h \left( X_b^B - x \right),
\]

as \( j = 0, 1, 2 \).

The local linear estimator (12), if compared to the Nadaraya–Watson estimator (11), does have several advantages. First of all, it does not suffer of the so-called “boundary effect” (Fan and Gijbels, 1996), consisting in being severely inefficient when \( x \) is close at the extremes of its range. Secondly, since it is based on a first-order local fit, it does not really need to assume that the variance \( V[\epsilon | X = x] \) is independent of \( x \), because it is approximately the same in a local neighbourhood of \( x \).

A crucial element, in determining the performance of the local polynomial estimator, is the choice of the bandwidth \( h \). This point will be discussed in the simulation study of Section 5.

4. Evaluation of the matching noise for the imputation procedure based on local polynomial regression estimator

One of the key issues in order to assess the accuracy of imputation procedures is to study the discrepancy between the distribution that generates genuine data (i.e. the distribution of \((X, Z)\)) and the distribution that generates imputed data (i.e. the distribution of \((\tilde{X}, \tilde{Z})\)). For all the imputation procedures described in Section 3, based on donors selected according to a distance with the recipient \( x_\nu^A \), and from the independence of different observations, it turns out that the distribution of \((\tilde{X}, \tilde{Z})\) is given by:

\[
f_{\tilde{X}^A_{\tilde{Z}^A}, \tilde{Z}^A}(x, z) = \int f_{\tilde{X}^A_{\tilde{Z}^A}(x, t, z)} dt = f_X(x) \int f_{X^A_{X^A}(x|t)}(t|x) f_{Z^A_{Z^A|t}}(t) dt,
\]

where \( t \) is a vector of dimension \( k \geq 1 \). As a matter of fact, if \( X \) is categorical, \( A \) and \( B \) observe all the categories of \( X \), and distance hot deck is considered, the matching noise is null. Generally speaking, a continuous \( X \) does not allow the definition of an imputation procedure with a null matching noise. The matching noise will depend on two elements:

- the distance between the recipient \( x_\nu^A \) and the donors \( x_{B(j)}^B \);
- how \( \tilde{Z} \) depends on the observed nearest records \( x_{B(j)}^B \).
In Marella et al. (2008) the matching noise that affects kNN method is determined. It is proved that \( X_{b_i} | X_a \) converges in distribution to a \( k \)-dimensional vector whose elements are equal to \( x_0 \). Hence, stochastic kNN (9), distance hot deck (3) and selection of a random element from the \( k \) nearest neighbours (2) tend asymptotically to be matching noise free, while deterministic kNN (8) is unavoidably biased.

When imputing on the basis of local linear regression estimators, a similar result holds. Roughly speaking, since, as \( n_B \) increases,

(i) the estimated regression function \( \hat{m} \) becomes closer and closer to the population regression function;

(ii) the empirical distribution of the residuals \( \hat{z}_a \) tends to be closer and closer to the distribution of the (population) errors \( \epsilon_a \) then the distribution of \( Z_a \) becomes closer and closer to the distribution of \( Z_a \). In other words, the technique described in Section 3.2.2 is asymptotically matching noise free. This claim is rigorously stated in Propositions 1 and 2, whose proofs are deferred to the Appendix.

**Proposition 1.** Assume that the model (4) holds, with \((X_b, Z_b)\), \(b = 1, \ldots, n_B \) i.i.d. random variables. Assume further that \( \hat{m}(-) \) tends in probability to \( m(-) \) as \( n_B \) goes to infinity, and that \( F_{\epsilon}(x) = \Pr(\epsilon \leq x) \) is continuous. If

\[
\hat{F}_{n_B}(x) = \frac{1}{n_B} \sum_{b=1}^{n_B} I(\hat{\epsilon}_b \leq x),
\]

is the empirical distribution function (e.d.f.) based on the residuals \( \hat{\epsilon}_b \)'s, then

\[
\sup_x |\hat{F}_{n_B}(x) - F_{\epsilon}(x)|,
\]

converges in probability to zero as \( n_B \) goes to infinity.

**Proof.** See Appendix. \( \square \)

Proposition 1 shows that the Kolmogorov–Smirnov (KS) distance between \( \hat{F}_{n_B}(-) \) and \( F_{\epsilon}(-) \) tends to zero as \( n_B \) increases. This suggests using KS distance as a “natural” distance for measuring the matching noise in the sequel of the paper.

As a consequence of Proposition 1, it is now easy to conclude that matching based on local linear regression estimator is asymptotically “noise-free” (Proposition 2).

**Proposition 2.** Under the same assumptions of Proposition 1

\[
z^A_a = \hat{m}^B(x^A_a) + \epsilon^B,
\]

possesses, as \( n_B \) goes to infinity, the same distribution as

\[
z^A_a = m(x^A_a) + \epsilon.
\]

**Proof.** See Appendix. \( \square \)

### 5. A simulation study

In this section we perform a simulation experiment to evaluate the matching noise produced by the nonparametric imputation techniques described in Section 3. It is necessary to resort to simulation procedures because it is not always possible to compute explicitly the matching noise associated to a given imputation technique for finite sample sizes. The simulation study has been carried out by using the software R (R Development Core Team, 2004).

#### 5.1. The simulation framework

In more detail, 500 i.i.d. records from a normal distribution \( X \) with mean 1 and variance 5 have been generated. Four regression functions, plotted in Fig. 1 and listed below in Eqs. (17)–(20), have been used to model the relationship between the predictor \( X \) and the response variable \( Z \).

\[
m_1(x) = 0.4 \left( \frac{x + 5.7}{13.4} \right) + 1,
\]

\[
m_2(x) = 0.3 + 4 \left( \frac{x + 5.7}{13.4} \right) - 3 \left( \frac{x + 5.7}{13.4} \right)^2,
\]

\[
m_3(x) = 4 \left( \frac{x + 5.7}{13.4} \right) - 2 + 2 \exp \left\{ -16 \left[ 4 \left( \frac{x + 5.7}{13.4} \right) - 2 \right]^2 \right\}
\]

The function $m_1(x)$ is linear in $x$, while the second and fourth functions are quadratic and quartic functions respectively. The third function is bump shaped.

The four regression functions have an increasing degree of complexity, ranging from a simple linear function to a complicate quartic function. They are frequently used in literature in both simulation and applied studies. More precisely, $m_1(\cdot)$, $m_3(\cdot)$ are used in Fan and Gijbels (1996) whilst $m_4(\cdot)$ is used in the estimation of wage equation in labour economics (Black and Juhn, 2000).

Normal random errors have been used for all test functions, $\epsilon \sim N(0, \sigma^2)$ for different values of $\sigma^2$. More specifically, the values $\sigma^2 = (0.3)^2$, $\sigma^2 = (0.5)^2$ have been used.

Let the recipient file $A$ consist of these 500 observations, with $Z$ dropped. For each regression function $m_i(x)$, $i = 1, \ldots, 4$, the simulation involves the following steps:

1. A donor sample $B$ composed by $n_B$ i.i.d. records has been generated exactly as $A$, except that the $Z$ values are not dropped. Different values of $n_B = 800–2000(400)$ have been used.
2. The missing $Z$s have been imputed by the imputation techniques described in Section 3.
3. Steps 1 to 2 have been repeated 400 times.

In order to evaluate the closeness between the data generating model and the imputation generating model, a divergence measure based on the Kolmogorov–Smirnov distance (KS) has been used. At first, the matching noise for the marginal distribution of $Z$ has been evaluated. Formally speaking, for each donor sample $v$ (for $v = 1, 2, \ldots, 400$), KS distance compares the empirical distribution function (e.d.f.) of imputed values $\tilde{Z}$ in $A(\hat{F}_{Z,v}(z))$ with the e.d.f. of true values $(\hat{F}_{Z,A}(z))$. A mean of such values over the 400 donor files is then taken as a global divergence measure, namely:

$$KS_Z = \frac{1}{400} \sum_{v=1}^{400} KS_Z(v) = \frac{1}{400} \sum_{v=1}^{400} \left[ \sup_{-\infty < z < \infty} |\hat{F}_{Z,A}(z) - \hat{F}_{Z,v}(z)| \right].$$

In order to get information on the closeness of the two distributions $f_{X,Z}(x, z)$ (the distribution generating the imputed data) and $f_{X,Z}(x, z)$ (the distribution generating genuine data), the KS distance has been computed between the conditional

![Fig. 1. Plots of the regression functions $m_i(x)$, $i = 1, \ldots, 4$.](image-url)
distribution of $Z \mid X = x_i^a, a = 1, \ldots, 500 (F_{Z \mid X}(z))$ and the conditional empirical distribution $\hat{F}_{Z \mid X}(z)$. To get a synthetic measure, the average over the $n_b = 500$ values has been computed:

$$E[KS^2_2] \approx \frac{1}{500} \sum_{a=1}^{500} KS_2(x_i^a) = \frac{1}{500} \sum_{a=1}^{500} \left[ \sup_{-\infty < z < \infty} |F_{Z \mid X}(z) - \hat{F}_{Z \mid X}(z)| \right].$$

Both kNN and local linear estimators require the specification of smoothing parameters ($k$ for the kNN method and $h$ for the local linear estimator, respectively).

According to the variance-bias trade off, $k$ has to be defined as a function of sample size $n_b$ such that $k/n_b \to 0$, as $n_b \to \infty$. The value of $k = \sqrt{n_b}$ has been chosen, as in Silverman (1986, p. 19), where such a value is recommended as an approximation of the asymptotically optimal value of $k$.

As shown in Section 3.2.2, the local linear regression estimator $\hat{m}(x)$ is obtained by fitting local straight lines in a neighborhood of $x$, with weights given by a kernel function $K$. In the simulation we have used a Gaussian kernel. An important point is the bandwidth selection, whose magnitude influences the amount of local smoothing. In the sequel we consider three selection rules of the smoothing parameter:

- The “Rule of Thumb” (Rot) bandwidth selection;
- The “Generalized Cross Validation” (Gcv) bandwidth selection;
- The bandwidth selection rule (Plug) given by Ruppert et al. (1995);

Rot is a crude bandwidth selector but requires little programming effort. Besides it is so little time consuming that other methods are hard to compete with. Essentially, the Rot bandwidth estimator estimates the unknown quantities $(\sigma_i^2, m_i^\circ)$, $i = 1, \ldots, 4$ appearing in the asymptotically optimal constant bandwidth fitting a polynomial of order 4 to $m_i(x), i = 1, \ldots, 4$ (Härdle, 1990).

Gcv is a simplification of the ordinary cross-validation bandwidth selection rule having the advantage to be less computationally intensive, since it does not require to fit the model $n$ times, one for each delete-one data (Loader, 2004). In the Gcv procedure the prediction error is estimated in a grid of points defined as $h_j = Ch_{j-1}, j \geq 1$ with $C = 0.1$. We start from $h = h_0 = h_{\min}$ and we stop when $h > h_{\max}$, where $h_{\min} = (\max x - \min x)/n_b$ and $h_{\max} = (\max x - \min x)/2$.

Plug is an adaptation of a plug-in bandwidth selector, see Ruppert et al. (1995). The basic idea is to estimate the unknown quantity $(\sigma_i^2, m_i^\circ)$ in the asymptotically optimal constant bandwidth by partitioning the range of $X$ into $n_b$ blocks, and by fitting a quartic polynomial in each block. The number $N$ is chosen by Mallows’s approach (Mallows, 1973) by the set 

$$C_p(N) = RSS(N)/(RSS(N_{\max})/(n - 5N_{\max})) - (n - 10N).$$

where $RSS(N)$ is the residual sum of squares based on a “blocked quartic fit” over $n_b$ blocks. Clearly if there are many oscillations in the regression function, then higher values of $N$ could be considered. Recommendations for $N_{\max}$ are in Ruppert et al. (1995).

5.2. Simulation results

We show here the simulation results for the case $\sigma^2 = (0.3)^2$. Similar results also hold in the case $\sigma^2 = (0.5)^2$.

In Fig. 2 we report for each regression function the matching noise produced by distance hot deck, random kNN, mean kNN plus random residual and by the deterministic imputation method (7) based on the local linear regression estimator of $m_i(x), i = 1, \ldots, 4$, as described in Section 3.2.2.

Let LRot, LGcv, LPlug be the deterministic imputation methods (7) based on the local linear regression estimators (12) coming from the bandwidth selectors mentioned above. It can be seen that the magnitude of the matching noise for all these different imputation procedures is essentially the same. These imputation techniques underestimate variability, since the replacement of each missing item by the expected value of $k$ nearest neighbors implies that the synthetic distribution of $Z \mid X$ is concentrated on the expected value of $Z \mid X$. The mean kNN plus random residual seems to perform better among all regression function estimators. Fig. 2 also suggests that the mean kNN plus random residual works better when the population regression function is “complex”.

The significance of the differences between matching noises associated to mean kNN plus random residual, distance hot deck and random kNN respectively has been tested by the usual T-test for the difference of means. In Tables 1 and 2 we report the test statistic (denoted by $\tau(m_i), i = 1, \ldots, 4$) for each regression function and for different values of $n_b$.

As a matter of fact, as $n_b$ increases the T-test statistical values decrease towards the test acceptance region for the four scenarios $m_i, i = 1, \ldots, 4$ (significance level 0.05). More precisely, for donor file sizes large enough and for simple regression functions (linear, quadratic) the matching noise differences between random kNN and mean kNN plus random residual are not significant. As a consequence, the performance of the imputation techniques depends on both data generating model complexity and donor file size. However, for $n_b$ large enough the methods have a similar behavior in reconstructing the marginal distribution of $Z$.

The methods LRot, LGcv, LPlug are deterministic and do not improve the mean kNN plus random residual performance. In order to recover a part of the data variability, their stochastic version (6) has been considered, where the residual is drawn at random from the residuals distribution obtained through the implementation of the same method on the donor file $B$. The results are reported in Fig. 3.
For all test functions, the stochastic imputation techniques based on both kNN estimator and local linear estimator seem to perform better than their deterministic counterparts: adjusting the regressed values in order to account for the residual variability reduces the matching noise. Such a behaviour is more evident for complex regression functions (i.e. functions 3 and 4). The differences between such imputation methods checked by performing the usual difference of means T-test are not significant for $n_B \geq 1200$ and $p$-value $= 0.01$. As a consequence, when comparing the marginal distribution of $Z$, the preference will be given to the mean kNN plus random residual since it is computationally easier and does not require any bandwidth selection.

In Fig. 4 we report the discrepancy measure (22) for all the stochastic nonparametric imputation techniques described in Section 3.

For each test function distance hot deck and LRot plus random residual seem to be the best methods. For the regression functions $m_i(x), i = 2, 3, 4$, LGcv method actually performs as good as LRot and distance hot deck methods. In all cases, the corresponding $E[KS^2]$ are not significantly different for $n_B \geq 800$ and $p − value = 0.03$. 

---

**Please cite this article in press as:** Conti, P.L., et al., Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.07.041
As previously stressed, (22) is a crude measure of divergence between the conditional distribution of $Z \mid X = x_a$, $F_{Z \mid x_a}(z)$, and the conditional empirical distribution $\hat{F}_{Z \mid x_a}(z)$. In order to evaluate the influence of the location point $x_a$ on the performance of the imputation techniques, the KS distance $K_S(x_a)$ has been evaluated for different values of $x_a$. As known, the boundary effect is more evident for the kNN than for the local linear estimator of the regression function. When $x_a$ is close to the boundaries, the kNN estimator is based on the computation of averages in an asymmetric region of $x_a$, consisting of a fixed number of $k$ points. Hence, the matching noise of the stochastic imputation method based on the kNN estimator could be severely high when $x_a$ is close at the extremes of its observational range. Besides, when the regression function is very steep at the boundaries the kNN estimator is expected to be more biased (for instance, $m_4(x)$), and hence the corresponding imputation method is affected by a severe matching noise. Such results are not reported to save space.

As a matter of fact, when comparing the conditional distribution of $Z \mid X$ a slight preference could be given to distance hot deck since it is more easily implemented and surely less computationally intensive. The same results have been obtained assuming that errors have common variance $\sigma^2 = 0.5^2$.

6. Conclusions

In this paper, a method of imputation based on the local linear estimation of the regression function of the variables of interest has been introduced and compared (in terms of matching noise) to other popular imputation techniques (hot deck methods and methods based on kNN estimators).

On the theoretical ground imputation based on local linear regression is asymptotically matching noise free. Comparisons made by simulation show that the higher the complexity of the functional relationship between the predictor $X$ and the response variable $Z$, the better the performance of the imputation method based on the local linear regression estimator.

The performance of imputation based on the local linear regression estimator is close to that of mean kNN plus random residual for the reconstruction of the marginal distribution of $Z$, and to that of the distance hot deck when the interest is in the conditional distribution of $Z \mid X$. As a result, this method offers an advantageous compromise for a good preservation of both the marginal and conditional distributions.

As far as the bandwidth selection is concerned, LRot and LGcv generally give good results. The LGcv method performs better when the population regression function is complex, far from linearity. This result parallels analogous results obtained

Fig. 4. $E[KS^2]$ for distance hot deck, random kNN, mean kNN + residual, LRot + residual, LGcv + residual, LPlug + residual.

by Marron and Wand (1992) for nonparametric density estimation. In that case, cross validation gives good results when the density function to be estimated is particularly rough.

**Acknowledgements**

Thanks are due to two anonymous referees for their helpful comments.

**Appendix**

**Proof of Proposition 1.** Let $F_{nB}(x)$ be the empirical distribution function based on the errors $\epsilon_{b}$:

$$F_{nB}(x) = \frac{1}{n_B} \sum_{b=1}^{n_B} I(\epsilon_b \leq x).$$  \hspace{1cm} (24)

We first show that

$$|\hat{F}_{nB}(x) - F_{nB}(x)|,$$

converges in probability to zero pointwise as $n_B$ goes to infinity. First of all, it is not difficult to see that

$$E[|\hat{F}_{nB}(x) - F_{nB}(x)|] = E\left[\frac{1}{n_B} \sum_{b=1}^{n_B} |I(\hat{\epsilon}_b \leq x) - I(\epsilon_b \leq x)| \right] \leq \frac{1}{n_B} \sum_{b=1}^{n_B} E[|I(\hat{\epsilon}_b \leq x) - I(\epsilon_b \leq x)|]$$

$$= \frac{1}{n_B} \sum_{b=1}^{n_B} \Pr(|I(\hat{\epsilon}_b \leq x) - I(\epsilon_b \leq x)| = 1) = 1.$$
\begin{equation}
\frac{1}{n_B} \sum_{b=1}^{n_B} \left[ \Pr(\hat{e}_b \leq x, e_b > x) + \Pr(\hat{e}_b > x, e_b \leq x) \right],
\end{equation}

where \(|I(\hat{e}_b \leq x) - I(e_b \leq x)| = 1\) if either \((e_b \leq x, \hat{e}_b > x)\) or \((e_b > x, \hat{e}_b \leq x)\), and \(|I(\hat{e}_b \leq x) - I(e_b \leq x)| = 0\) otherwise. Next, we may write

\[
\Pr(\hat{e}_b \leq x, e_b > x) = \Pr(z_b - \hat{m}(x_b) + x, z_b > m(x_b) + x) \\
= \Pr(m(x_b) + x < z_b - \hat{m}(x_b) + x) \\
= \mathbb{E}_b[\Pr(m(x_b) + x < z_b - \hat{m}(x_b) + x, |\hat{m}(x_b) - m(x)| < \delta|x_b|)] \\
+ \mathbb{E}_b[\Pr(m(x_b) + x < z_b - \hat{m}(x_b) + x, |\hat{m}(x_b) - m(x)| \geq \delta|x_b|)] \\
\leq \mathbb{E}_b[\Pr(m(x_b) + x < z_b - \hat{m}(x_b) + x, |\hat{m}(x_b) - m(x)| < \delta|x_b|)] \\
+ \mathbb{E}_b[\Pr(|\hat{m}(x_b) - m(x)| \geq \delta|x_b|)],
\]

for every \(\delta > 0\). Since \(\hat{m}(\cdot)\) is a consistent estimator of \(m(\cdot)\), the second expected value in (27) goes to zero as \(n_B\) goes to infinity. As far as the first one is concerned, from the inequality

\[
\mathbb{E}_b[\Pr(m(x_b) + x < z_b - \hat{m}(x_b) + x, |\hat{m}(x_b) - m(x)| < \delta|x_b|)] \\
= \mathbb{E}_b[\Pr(m(x_b) + x < z_b - \hat{m}(x_b) + x, m(x_b) - \delta < m(x_b) < m(x_b) + \delta|x_b|)] \\
\leq \mathbb{E}_b[\Pr(m(x_b) + x < z_b < m(x_b) + x + \delta|x_b|)] \\
= \Pr(x < e_b \leq x + \delta) < \tau,
\]

and from the continuity of \(F_1(x)\), it is seen that for every \(\tau > 0\) there exists \(n_{B,\tau}\) “large enough” such that (28) is smaller than \(\tau\) for any \(n_B \geq n_{B,\tau}\). The same consideration holds for \(\Pr(\hat{e}_b > x, e_b \leq x)\) in (26). This clearly implies that, for each fixed \(x\), the quantity

\[
\sqrt{n_B} \left| \hat{F}_{n_B}(x) - F_{n_B}(x) \right|
\]

converges in probability to zero as \(n_B\) goes to infinity. From the Glivenko–Cantelli theorem we know that

\[
\sup_x \left| \hat{F}_{n_B}(x) - F_{n_B}(x) \right|
\]

converges almost surely to zero as \(n_B\) goes to infinity. Since \(F_{n_B}(x)\) is continuous, this is enough to conclude that (14) holds. That is, from (29) and (30) it is immediate to see that

\[
\sup_x \left| \hat{F}_{n_B}(x) - F_{n_B}(x) \right|
\]

converges in probability to zero as \(n_B\) goes to infinity, i.e. the e.d.f. of residuals tends to reproduce the d.f. of \(e_b\). \(\Box\)

**Proof of Proposition 2.** Immediate consequence of Proposition 1. \(\Box\)

**References**


Please cite this article in press as: Conti, P.L. et al., Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.07.041
Yates, F., 1933. The analysis of replicated experiments when the field results are incomplete. Emporium Journal of Experimental Agriculture 1, 129–142.