Computational Statistics and Data Analysis 53 (2008) 354–365

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

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

Pier Luigi Conti [a], Daniela Marella [a,*], Mauro Scanu [b]

[a] *Università di Roma "La Sapienza", Italy*
[b] *Istituto Nazionale di Statistica, Italy*

ARTICLE INFO

Article history:
Received 14 June 2007
Received in revised form 24 July 2008
Accepted 24 July 2008
Available online 8 August 2008

ABSTRACT

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.

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.

* 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.
doi:10.1016/j.csda.2008.07.041

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,

$$\begin{cases} \mathbf{x}^A = \left(x_1^A, \ldots, x_{n_A}^A\right), \\ \left(\mathbf{x}^B, \mathbf{z}^B\right) = \left((x_1^B, z_1^B), \ldots, (x_{n_B}^B, z_{n_B}^B)\right), \end{cases} \tag{1}$$

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, *i.e.* missing Z values are replaced by actually observed values chosen appropriately among the n_B complete records in B. Usually, donor values are selected according to a distance between observed and incomplete records on X (Aluja-Banet et al., 2007). Two of the most popular procedures are distance and random hot deck imputation. Hot deck methods have been largely studied in the statistical literature, see Kalton and Kasprzyk (1986) and Little and Rubin (1987). By far, distance hot deck is the most used. Generally speaking, hot deck methods possess properties that are frequently considered as attractive by users: (i) they are nonparametric, because they do not need any explicit definition of a parametric data generation model; (ii) they impute "live values", *i.e.* actually observed values; (iii) they are able to reproduce the marginal and conditional distributions of the variable to impute quite well (at least for large samples). In our opinion property (iii) is the most important one, since it guarantees that the larger n_B, the smaller the matching noise.

As already said, the comparison among different imputation procedures is based on their matching noise, which is essentially the discrepancy (measured by the Kolmogorov–Smirnov distance) between the distribution generating genuine data and the distribution generating imputed data (Paass, 1985). If these two distributions coincide, the imputed data set can be analyzed as if it was a completely observed data set generated by the distribution generating genuine data (the joint distribution of (X, Z)). Otherwise estimators based on the complete synthetic data set could be inappropriate for inferring properties of the model underlying data. An example of study of the matching noise for a class of nonparametric imputation procedures based on kNN methods (including distance hot deck) is in Marella et al. (2008), where $f(x, z)$ was assumed to be a bivariate normal density. This assumption is dropped in the present paper.

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 $(X, \tilde{Z})$ 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 $\mathbf{b}(a) = (b_1(a), \ldots, b_k(a))$ be the labels of the $k \geq 1$ nearest neighbours of x_a^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

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

where $d(.,.)$ is the Euclidean distance. Let $\mathbf{x}_{\mathbf{b}(a)}^B = (x_{b_1(a)}^B, x_{b_2(a)}^B, \ldots, x_{b_k(a)}^B)$ and $\mathbf{z}_{\mathbf{b}(a)}^B = (z_{b_1(a)}^B, z_{b_2(a)}^B, \ldots, z_{b_k(a)}^B)$ 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, $x_{\mathbf{b}(a)}^B$, 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

$$\tilde{z}_a^A = z_{\tilde{b}(a)}^B, \quad a = 1, \ldots, n_A. \tag{2}$$