

Nonparametric approaches to statistical matching

Approcci non parametrici al problema dell'abbinamento statistico

Pier Luigi Conti¹, Daniela Marella², Mauro Scanu³

¹ DSPSA, Università di Roma “La Sapienza”

² daniela.marella@yahoo.it

³ Dipartimento della Produzione Statistica, ISTAT, Roma

Riassunto: In questo lavoro si studia il problema della discrepanza fra la distribuzione generatrice dei dati imputati e quella che genera i dati veri, non osservati (matching noise). In particolare, si confrontano diverse tecniche di imputazione non parametrica, sia dal punto di vista teorico, sia mediante un’esperimento di simulazione.

Keywords: Imputation, matching noise, kNN method, hot deck.

1. Introduction and notation

Imputation of missing items is a usual practice in applied statistics. One of the most common approaches is *hot deck*. It consists in properly selecting donors from the respondents of a data set, and imputing the donor values to the unobserved ones. This approach aims at obtaining a completed, synthetic, data set that is much easier to analyze. A natural question is: To which extent is the completed data set reliable? In fact, it can be actually used for any kind of analysis if the mechanism that generates imputations coincides with the random mechanism that generates “true” observations. The discrepancy between these two processes is the *matching noise* (cfr. Paass, 1985).

This paper aims at studying the matching noise of the distance hot deck procedure, as well as of some nonparametric alternatives, in the special case of statistical matching (cfr. Rässler, 2002). Let A, B be two samples of size n_A and n_B , respectively, of independent and identically distributed (*i.i.d.*) records generated from a $P + Q$ dimensional random variable (r.v.) $(\mathbf{X}, \mathbf{Z})$ with joint density function (d.f.) $f(\mathbf{x}, \mathbf{z})$. From now on, we will denote by $\mathbf{T}_c^C$ the c th observation of the variate T in the sample C ($C = A, B$, $c = 1, \dots, n_C$, $\mathbf{T} = \mathbf{X}, \mathbf{Z}$). The r.v. $\mathbf{Z}$ is not observed in A , and is imputed using B as a set of donors *via* hot deck procedures. Distance hot deck consists in selecting, for each $a = 1, \dots, n_A$, the donor $b_1(a) \in B$ such that $d(\mathbf{x}_a^A, \mathbf{x}_{b_1(a)}^B) = \min_{b \in B} d(\mathbf{x}_a^A, \mathbf{x}_b^B)$, where $d(\cdot, \cdot)$ is a distance function. In the sequel we will confine ourselves to Euclidean distances $d(\mathbf{x}_a^A, \mathbf{x}_b^B) = \{(\mathbf{x}_b^B - \mathbf{x}_a^A)' D (\mathbf{x}_b^B - \mathbf{x}_a^A)\}^{1/2}$, D being a positive definite matrix. The completed A consists of the records $(\mathbf{x}_a^A, \mathbf{z}_{b_1(a)}^B)$. As remarked in Chen and Shao (2001), distance hot deck provides asymptotically unbiased and consistent estimators for population means as well as for quantiles. No attempt to define the correspondent matching noise has been made. Distance hot deck is nonparametric because it does not require any specific parametric assumption on the d.f. of $(\mathbf{X}, \mathbf{Z})$. Furthermore, it corresponds to use the k nearest neighbour (k NN) method with $k = 1$. It can be easily generalized for $k > 1$.

2. Evaluation of matching noise

Let $\mathbf{b}(a)$ be the vector of the labels $b_1(a), \dots, b_{n_B}(a)$ corresponding to the ordered nearest neighbours of $\mathbf{x}_1^B, \dots, \mathbf{x}_{n_B}^B$ to $\mathbf{x}_a^A$, $a = 1, \dots, n_A$. Let $\tilde{\mathbf{Z}} = g(\mathbf{Z}_{\mathbf{b}(a)}^B)$ be the r.v. representing the imputed values of $\mathbf{Z}$ in A . We aim at evaluating the d.f. of the r.v. $(\mathbf{X}, \tilde{\mathbf{Z}})$, i.e. the distribution of the completed records in A , and its discrepancy w.r.t. $f(\mathbf{x}, \mathbf{z})$. Since $\tilde{\mathbf{Z}}$ is assigned to each record $a \in A$ by computing the distance between $\mathbf{X} = \mathbf{x}_a^A$ and the observed $\mathbf{X}$ values in B , we have the relationship

$$f_{\mathbf{X}_a^A, \mathbf{X}_{\mathbf{b}(a)}^B, \tilde{\mathbf{Z}}}(\mathbf{x}, \mathbf{t}, \tilde{\mathbf{z}}) = f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{X}_{\mathbf{b}(a)}^B | \mathbf{X}_a^A}(\mathbf{t} | \mathbf{x}) f_{\tilde{\mathbf{Z}} | \mathbf{X}_{\mathbf{b}(a)}^B}(\tilde{\mathbf{z}} | \mathbf{t}), \quad (1)$$

i.e. the imputed value is independent of the recipient $\mathbf{X}$, given the observed $\mathbf{X}$ in B .

From equation (1), the joint d.f. $f(\mathbf{x}, \tilde{\mathbf{z}})$ for distance hot deck, i.e. for $g(\mathbf{Z}_{\mathbf{b}(a)}^B) = \mathbf{z}_{b_1(a)}^B$, is given by $\int f_{\mathbf{X}_a^A}(\mathbf{x}) f_{\mathbf{X}_{b_1(a)}^B | \mathbf{X}_a^A}(\mathbf{t} | \mathbf{x}) f_{\tilde{\mathbf{Z}} | \mathbf{X}_{b_1(a)}^B}(\tilde{\mathbf{z}} | \mathbf{t}) d\mathbf{t}$, where $f_{\mathbf{X}_{b_1(a)}^B | \mathbf{X}_{b_1(a)}^A}(\mathbf{t} | \mathbf{x})$ is the (conditional) d.f. of the nearest neighbour $\mathbf{X}_{b_1(a)}^B$ of $\mathbf{X}_a^A$. The matching noise for hot deck distance is obtained in Proposition 1 (a proof is in: Conti, Scanu, 2005)

Proposition 1 *The conditional density of $\mathbf{X}_{b_1(a)}^B$, given $\mathbf{X}_a^A = \mathbf{x}_a^A$, is equal to*

$$f_{\mathbf{X}_{b_1(a)}^B | \mathbf{X}_a^A}(\mathbf{x}) = f_{\Psi_{n_B:1}}(\mathbf{x} - \mathbf{x}_a^A), \quad (2)$$

where $f_{\Psi_{n_B:1}}(\psi) = P(W \geq \psi' D \psi)^{n_B-1} f_{\Psi}(\psi)$, $\Psi_b = \mathbf{X}_b^B - \mathbf{X}_a^A$, $\Psi_{n_B:1} \leq \dots \leq \Psi_{n_B:n_B}$ are the ordered values Ψ_b s, $W = \psi' D \psi$, and f_{Ψ} is the conditional d.f. of Ψ_b given $\mathbf{X}_a^A = \mathbf{x}_a^A$.

Proposition 1 states that the matching noise is absent when (2) is concentrated in $\mathbf{x}_a^A$. This is impossible when $\mathbf{X}$ is a continuous r.v.. However, one expects that matching noise decreases as n_B increases. This is formally stated in Proposition 2 (Conti, Scanu, 2005).

Proposition 2 *Let ϵ be a positive number, and let ϵ be a P -dimensional vector with components all equal to ϵ . Denote further by $(\mathbf{x}_a^A - \epsilon, \mathbf{x}_a^A + \epsilon)$ the open square of sides 2ϵ centered around $\mathbf{x}_a^A$, and assume that the support of W contains an interval of the form $(0, t)$ for some positive t . Then:*

$$\lim_{n_B \rightarrow \infty} P(\mathbf{X}_{b_1(a)}^B \notin (\mathbf{x}_a^A - \epsilon, \mathbf{x}_a^A + \epsilon) | \mathbf{X}_a^A = \mathbf{x}_a^A) = 0 \quad \forall \epsilon > 0.$$

The above results are easily extended to the case of k NN imputation, with $k > 1$. Given $\mathbf{x}_a^A$, its k nearest neighbours are the k observations $\mathbf{x}_{b_1(a)}^B, \dots, \mathbf{x}_{b_k(a)}^B$ in B which are closest to $\mathbf{x}_a^A$, according to the distance d . The imputed value $\tilde{\mathbf{Z}}_a^A$ is then a function $g(\mathbf{Z}_{b_1(a)}^B, \dots, \mathbf{Z}_{b_k(a)}^B)$. Common examples are the arithmetic mean of $\mathbf{Z}_{b_j(a)}^B$ s, or their median, or a randomly chosen value from $\mathbf{Z}_{b_j(a)}^B$ s.

Proposition 3 *Under k NN imputation, and denoting by $\mathbf{b}_k(a)$ the vector of components (k NN labels) $b_1(a), \dots, b_k(a)$, the conditional d.f. of $\mathbf{X}_{\mathbf{b}_k(a)}^B$, given $\mathbf{X}_a^A = \mathbf{x}_a^A$, is equal to*

$$f_{\mathbf{X}_{\mathbf{b}_k(a)}^B | \mathbf{X}_a^A}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k) = P(W \geq \psi'_k D \psi_k)^{n_B-k} \prod_{j=1}^k f_{\Psi}(\psi_j)$$

where $\psi_j = \mathbf{x}_j - \mathbf{x}_a^A$, and, again, f_{Ψ} is the conditional d.f. of Ψ_b given $\mathbf{X}_a^A = \mathbf{x}_a^A$.

Similarly to the hot deck distance imputation, the following result holds.

Proposition 4 *Using the same notation as in Proposition 2, and writing $\mathbf{X}_{\mathbf{b}_k(a)}^B \in (\mathbf{x}_a^A - \epsilon, \mathbf{x}_a^A + \epsilon)$ if and only if $\mathbf{X}_{\mathbf{b}_j(a)}^B \in (\mathbf{x}_a^A - \epsilon, \mathbf{x}_a^A + \epsilon)$ for every $j = 1, \dots, k$, we have:*

$$\lim_{n_B \rightarrow \infty} P(\mathbf{X}_{\mathbf{b}_k(a)}^B \notin (\mathbf{x}_a^A - \epsilon, \mathbf{x}_a^A + \epsilon) | \mathbf{X}_a^A = \mathbf{x}_a^A) = 0 \quad \forall \epsilon > 0.$$

Despite its similarity with Proposition 2, there is a fundamental difference. Proposition 2 tells us that, if n_B is large enough, then the matching noise is negligible because $\mathbf{X}_{\mathbf{b}_1(a)}^B$ is “close” to $\mathbf{X} = \mathbf{x}_a^A$ with high probability, and hence the conditional distribution of $\mathbf{Z}$, given $\mathbf{X}_{\mathbf{b}_1(a)}^B$, is close to the conditional d.f. of $\mathbf{Z}$, given $\mathbf{X} = \mathbf{x}_a^A$. Proposition 4 implies that all k components of $\mathbf{X}_{\mathbf{b}_k(a)}^B$ are “close” to $\mathbf{X} = \mathbf{x}_a^A$. Hence, the conditional d.f. of $\mathbf{Z}_{\mathbf{b}_j(a)}^B$, given $\mathbf{X}_{\mathbf{b}_k(a)}^B$ is close to the conditional d.f. of $\mathbf{Z}$, given $\mathbf{X} = \mathbf{x}_a^A$, $j = 1, \dots, k$. This does not imply that the conditional d.f. of $g(\mathbf{Z}_{\mathbf{b}_1(a)}^B, \dots, \mathbf{Z}_{\mathbf{b}_k(a)}^B)$, given $\mathbf{X}_{\mathbf{b}_k(a)}^B$ is close to the d.f. of $\mathbf{Z}$ given $\mathbf{X}$. For instance, if g is the mean of $\mathbf{Z}_{\mathbf{b}_j(a)}^B$ s, $g(\mathbf{Z}_{\mathbf{b}_1(a)}^B, \dots, \mathbf{Z}_{\mathbf{b}_k(a)}^B)$ tends to the distribution of the sample mean of k i.i.d. copies of $\mathbf{Z}$, given $\mathbf{X}$.

An alternative imputation procedure, again based on k NN, could be the following. Assume a (multivariate) nonparametric regression model $\mathbf{Z} = h(\mathbf{X}) + \mathbf{U}$, with $E(\mathbf{U}|\mathbf{X}) = \mathbf{0}$, $E(\mathbf{U}\mathbf{U}'|\mathbf{X}) = \mathbf{G}$, $\mathbf{G}$ being a diagonal matrix independent of $\mathbf{X}$. A simple idea consists in (i) estimating first $h(\mathbf{X})$ by a nonparametric estimator $\hat{h}$; (ii) defining the residuals $\mathbf{e}_b^B = \mathbf{z}_b^B - \hat{h}(\mathbf{x}_b^B)$, $b = 1, \dots, n_B$; (iii) drawing at random a residual $\mathbf{e}_{b^*}^B$; (iv) imputing $\tilde{\mathbf{z}}_a^A = \hat{h}(\mathbf{x}_a^A) + \mathbf{e}_{b^*}^B$. If $\hat{h}(\cdot)$ is a consistent estimator of $h(\cdot)$ then it is not difficult to see that the matching noise of this procedure vanishes as n_B increases. In this way we have defined a (nonparametric) class of consistent imputation methods. In section 3, the performance of this imputation technique is compared to the hot deck one in the special case of $\hat{h} = k$ NN estimator of h . A more extensive study to assess the relative merits of such imputation techniques methods will be accomplished in a subsequent work.

3. Experimental results and conclusions

In order to compare the performance of the imputation techniques used in this paper we have performed a simulation experiment. In detail, we have randomly generated 1 000 i.i.d. records from a bivariate normal distribution (X, Z) with means 1, 3, and variances 5, 4, respectively, and covariance 3. Let A consist of the first 500 observations, with Z dropped, and B of the last 500 complete observations. The missing Z s have been imputed by the following techniques: (i) distance hot deck, with $d(x_a^A, x_b^B) = |x_a^A - x_b^B|$; (ii) k NN with different k s, $k > 1$, and $g(\cdot)$ corresponding to the mean function (mean k NN) and to a random draw (random k NN), respectively; (iii) k NN estimator of $h(X) = E(Z|X)$ plus random residual, as defined at the end of section 2.

Distance hot deck and random k NN of imputed values are, loosely speaking, almost equivalent when n_B is large enough. This is an obvious consequence of Propositions 2, 4 because under random k NN the k possible donors are approximately drawn from the same conditional d.f. of Z given X . Mean k NN underestimates variability and this worsens as k increases. However, it may better estimate the relationship (in expectation) between Z

Figure 1: Matching noise for distance hot deck (left), and k NN, $k = 2$, mean and random (right).

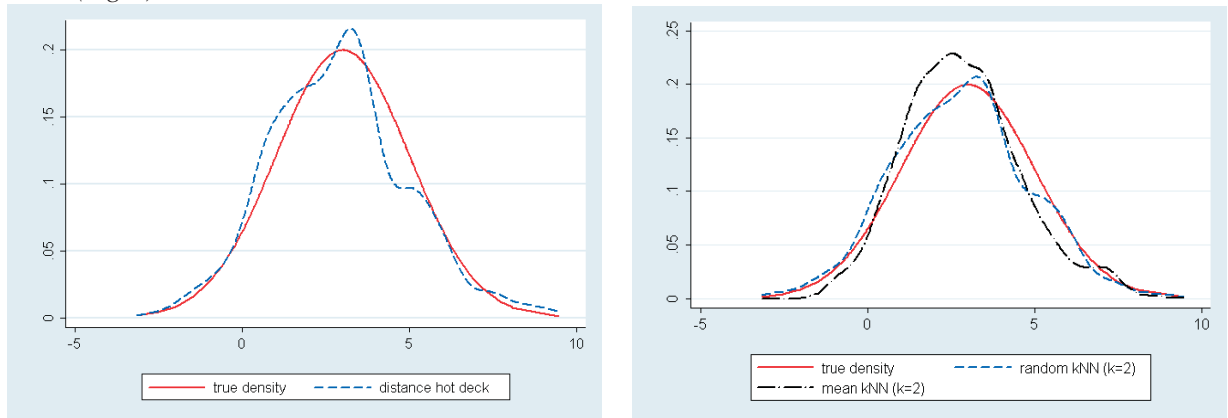
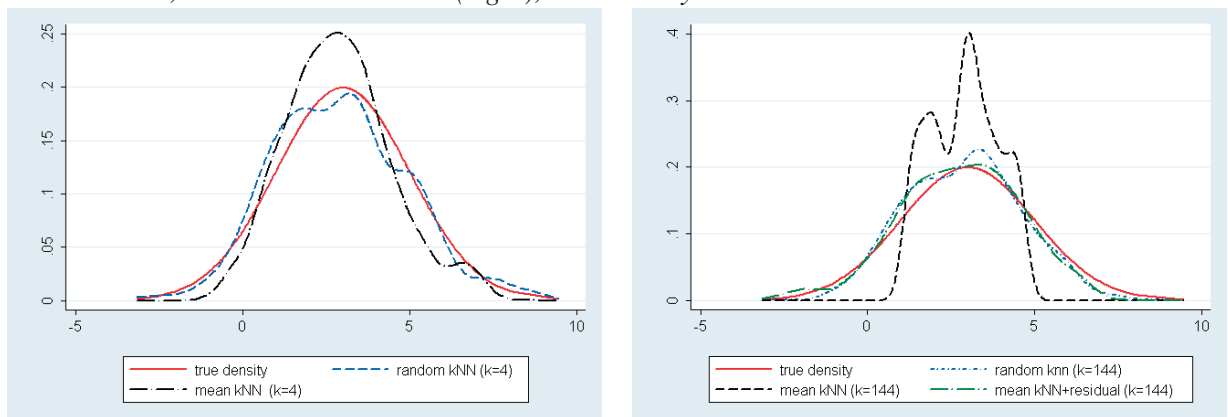


Figure 2: Matching noise for k NN method, $k = 4$, mean and random (left), and for k NN random, mean and $\hat{h}$ +residual (right), k chosen by cross validation.



and X . k NN estimator of $h(X)$ plus random residual seems to perform slightly better than hot deck. This remark stimulates further studies in this direction, e.g. by evaluating the speed at which the matching noise vanishes for different methods.

References

- Chen, J., Shao J. (2001) Jackknife Variance Estimation for nearest-Neighbor Imputation *JASA*, 96, 260-269.
- Conti P.L., Scanu M. (2005) On the evaluation of matching noise produced by nonparametric imputation techniques *Rapporto Tecnico 2005 n. 7, DSPSA, Università di Roma "La Sapienza"*.
- Paass, G. (1985) Statistical record linkage methodology, state of the art and future prospects. *Bulletin of the International Statistical Institute*, Proceedings of the 45th Session, vol. LI. Book 2.
- Rässler, S. (2002) *Statistical Matching: a Frequentist Theory, Practical Applications and Alternative Bayesian Approaches* Springer.