Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 5913–5917
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Convergence of random k-nearest-neighbour imputation

Fredrik A. Dahl

Helse Øst Health Services Research Centre, Akershus University Hospital, Mail drawer 95, NO-1478 Lørenskog, Norway

Received 7 July 2006; received in revised form 6 November 2006; accepted 6 November 2006
Available online 29 November 2006

Abstract

Random k-nearest-neighbour (RKNN) imputation is an established algorithm for filling in missing values in data sets. Assume that data are missing in a random way, so that missingness is independent of unobserved values (MAR), and assume there is a minimum positive probability of a response vector being complete. Then RKNN, with k equal to the square root of the sample size, asymptotically produces independent values with the correct probability distribution for the ones that are missing. An experiment illustrates two different distance functions for a synthetic data set.

Keywords: Hot deck; Imputation; Survey data; k-Nearest-neighbour; Convergence

1. Introduction

Our setting is this: we have a sample of multivariate, independent, identically distributed data, with missing values. We would like to fix the data set by reconstructing valid data values for the ones that are missing, a task commonly referred to as *imputation*. The most common application of imputation is with statistical survey data (Rubin, 2004), but it is also relevant in other fields, such as DNA micro array analysis (Troyanskaya et al., 2001). Applying imputation to a data set before analysing it may seem dubious from a scientific perspective, analysing computer-generated values as if they were real data. However, there have been computer simulation studies, e.g. Twisk and de Vente (2002) or Hawthorne and Elliott (2005) showing that, under suitable conditions, inference from imputed data sets is robust. Also, the alternative approach of deleting records with missing values may also introduce bias. Imputation is most appropriate when we plan to build complex statistical models with several independent variables. In this case the alternative of deleting all records where one of the values is missing may reduce the data set substantially, even though each univariate variable has few missing values.

Deterministic approaches to imputation treats each missing value as a problem of prediction, where the task is to reconstruct the most likely value for the ones that are missing. These include simple averaging procedures and more advanced model-based approaches like linear regression and EM-algorithms. In principle, any prediction method from statistical learning (Hastie et al., 2001) can be used for this purpose. Although deterministic imputation works reasonably well in many cases, it will tend to reinforce trends in the data set, resulting in falsely precise inference. To eliminate this bias, random imputation methods (Rubin, 2004) instead seek to predict the *conditional distribution* of the missing values, given the non-missing ones, and draw random values accordingly.

E-mail address: fredrik.dahl@ahus.no.

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.11.007

A popular class of imputation methods is so-called *hot deck* imputation (Reilly, 1993). For a given missing value, this algorithm searches the data set for records that are similar for the non-missing values. One of these records is chosen as a *donor*, and the donor's value is used for imputing the missing one. In random hot deck methods, the donor is chosen at random among a suitable set of donors. In the present paper we analyse a random hot deck algorithm that draws a donor from a set of k neighbours, defined by a metric. To our knowledge, this algorithm was first described by Little (1988).

In Section 2 we briefly review established definitions of how data can be missing. In Section 3 we prove asymptotic properties of the random k-nearest-neighbour (RKNN) algorithm under suitable conditions. Section 4 gives a discussion of the theoretical results, with emphasis on distance functions, and this issue is investigated further by an experiment in Section 5. Section 6 concludes the article.

2. Ways in which data can be missing

When imputing missing values, one has to make assumptions about their true distribution. The most favourable form of missingness is *missing completely at random* (MCAR), which means that the probability of a value missing is independent of all values in the data set, observed and unobserved. Missing at random (MAR) is less restrictive, as it allows the mechanism that produces missing values to depend on observed values, but not on unobserved ones. The most severe form of missingness is *missing not at random* (MNAR), which allows missingness to depend on missing values. If a data set is MNAR, one must include knowledge of the mechanism generating the missing values, in order to produce sound imputation. Note that it is impossible to test whether a data set is MNAR through the data set itself. These are established definitions, which have been attributed to Rubin.

3. RKNN imputation

Let $\{Z_i\}, i = 0, \ldots, n$ be a sequence of independent identically distributed random variables, of dimension m, with missing values. Note that the independence assumption states that Z_i is independent of $Z_{i'}$, when $i \neq i'$. The multidimensional distribution, which is common to all Z_i, may of course have dependent components. (In a statistical survey setting, the index i refers to a record of the data set, typically associated with a subject, and m gives the number of different questions, so that $Z_{i,j}$ represents subject i's response to question j.) Hence, different subject are assumed to give independent identically distributed responses, while a subject's responses to different questions may be dependent. For analytical clarity we assume that we start with a complete data set, where after missing values are deleted. We let $l < m$ be the number of missing values for Z_0, and without loss of generality we take these to be the first l components. The task is to impute these l values. For convenience, we split $Z_i = \binom{Y_i}{X_i}$, where $Y_i \in \mathbf{R}^l$ and $X_i \in \mathbf{R}^{m-l}$, so that Y_0 are the missing values of Z_0. We write $\mathbf{Y} = [Y_1 \ldots Y_n]$, $\mathbf{X} = [X_1 \ldots X_n]$, and $\mathbf{Z} = \binom{\mathbf{Y}}{\mathbf{X}}$. Note that the index 0 is excluded from these definitions. We let $d : \mathbf{R}^{2(m-l)} \to \mathbf{R}$ denote the metric, so that $d(X, X')$ gives the distance between X and X'. If X or X' has missing values, we let $d(X, X')$ assume the largest possible d-value one can get by completing X and X' with values from their domains. We otherwise assume that d is topologically equivalent to Euclidean distance.

Our version of the RKNN procedure can be formalized like this:

Let $k = \sqrt{n}$ (rounded upward).
For each $i \in \{1, \ldots, n\}$
    Let $d_i = d(X_0, X_i) + \mathrm{M}\chi$ (Y_i has a missing value).
Let $\eta \subset \{1, \ldots, n\}$ so that $|\eta| = k$ and $(i \in \eta, j \notin \eta) \Rightarrow d_i \leqslant d_j$.
Let I be a random element of η.
Let the imputed value be Y_I.

In the definition of d_i, χ is the indicator function, returning 1 if the argument is true, and 0 otherwise, while M is a (large or even infinite) penalty for Y_i having missing values. The computation of η must have some way of breaking ties, which may e.g. be a randomisation procedure.

In the following theorems we treat X_0 as given data, while Y_0, $\mathbf{X}$, and $\mathbf{Y}$ are random. We assume that the missing values are MAR, and assume that there exists a $p > 0$, so that $P(Z_i \text{ has no missing value}) > p$, for each i.

Theorem 1. $(Y_I|X_0) \underset{n\to\infty}{\to} (Y_0|X_0)$ *in distribution.*

Proof. The fraction of the points belonging to the k-neighbourhood, k/n, approaches zero. Because $P(Z_i$ has no missing value$) > p$ for each i, we have $\lim_{n\to\infty} \max_{i\in\eta}(d(X_0, X_i)) = 0$ in probability. (We notation-wise suppress the fact that the set η varies with n.) This implies that $X_I \underset{n\to\infty}{\to} X_0$ in distribution. Also, with probability 1, η will for sufficiently large n only contain indexes of points Z_i without missing values. Because the data set is MAR, the fact that the values of Y_0 are missing is independent of the observed values of X_I. It follows that when n increases, Y_I converges in distribution toward Y_0. □

The assumption that d be topologically equivalent to Euclidean distance is mild. For categorical data, this implies that the distance between two persons is zero if, and only if, their values are identical for all components.

In the next theorem we show that Y_I, the imputed value for Y_0, is close to independent of the rest of the data set, $\mathbf{Z}$. We do this by giving an upper bound on the correlation between Y_I and any measurable function f of $\mathbf{Z}$. (Readers unfamiliar with probability theory should not worry about the measurability condition, as this is a very mild assumption needed only to assure that $f(\mathbf{Z})$ is a random variable.)

Theorem 2. *Let* $f : \mathbf{R}^{n\times m} \to \mathbf{R}$ *be measurable, with* $Var(f(\mathbf{Z})) < \infty$ *and* $j \in \{1, \ldots, l\}$. *Then* $|cor(Y_{I,j}, f(\mathbf{Z}))| \leqslant n^{-1/4}$.

Proof. The function f that maximizes the correlation is the conditional expectation of $Y_{I,j}$ given $\mathbf{Z}$: $\operatorname{argmax}_{f(\mathbf{Z})}(cor(Y_{I,j}, f(\mathbf{Z})) = E(Y_{I,j}|\mathbf{Z}) = (1/k)\sum_{i\in\eta} Y_{i,j} = \bar{Y}_j$. (In order to make the proof more readable, we notation-wise suppress the index j for all Y-variables, from here on.) We decompose $Y_I = \sum_{i\in\eta} Y_i\chi_i$, where χ_i is an indicator variable of the event $I = i$. This gives

$$Cov(Y_I, \bar{Y}) = \frac{1}{k}\sum_{i,i'\in\eta} Cov(Y_i\chi_i, Y_{i'}) = \frac{1}{k}\sum_{i\in\eta} Cov(Y_i\chi_i, Y_i) = \frac{1}{k^2}\sum_{i\in\eta} Var(Y_i).$$

We also have $Var(\bar{Y}) = (1/k^2)\sum_{i\in\eta} Var(Y_i)$, and by a conditional variance argument

$$Var(Y_I) = \frac{1}{k}\sum_{i\in\eta}(Var(Y_i) + Var(E(Y_i))) \geqslant \frac{1}{k}\sum_{i\in\eta} Var(Y_i).$$

Piecing this together gives

$$cor(Y_I, \bar{Y}) = \frac{Cov(Y_I, \bar{Y})}{\sqrt{Var(Y_I)Var(\bar{Y})}} \leqslant \frac{1}{\sqrt{k}}.$$

Because we have $k \geqslant \sqrt{n}$, it follows that $cor(Y_I, f(\mathbf{Y})) \leqslant n^{-1/4}$.

The inequality $cor(Y_I, f(\mathbf{Y})) \geqslant -n^{-1/4}$ follows by repeating the argument after inverting the sign of f. □

Before the missing values were deleted, the data set was a sequence of independent identically distributed random variables. Our theorems state that the algorithm asymptotically produces data sets with the same properties of independence (Theorem 2) and identical distributions (Theorem 1).

4. Discussion

Concerning our choice of neighbourhood size function $k(n) = \sqrt{n}$, the proof of Theorem 1 works for any function $k(n)$ such that $k(n)/n$ tends to zero. The proof of Theorem 2, on the other hand, shows that $\lim_{n\to\infty}|cor(Y_I, f(\mathbf{Z}))| = 0$, provided $k(n)$ tends to infinity. Therefore, the favourable asymptotic properties hold for any neighbourhood size function of the form $k(n) = n^r$, where $r \in (0, 1)$. Our choice of $k(n) = \sqrt{n}$ is canonical in the sense of representing the mid-point of this interval.

Table 1
Estimated variances of the difference between actual and imputed values by sample size, for Euclidean and regression-based distance

	$n=2$	$n=4$	$n=8$	$n=16$	$n=32$	$n=64$	$n=128$	$n=256$	$n=512$	$n=1024$
d_E	14.00	9.14	5.02	4.54	3.68	3.36	3.04	2.86	2.70	2.60
d_R	14.00	8.37	3.96	3.51	2.70	2.47	2.27	2.18	2.10	2.07

Note that our assumption of a probability $p > 0$ of a data vector being complete cannot be discarded, as the MAR condition by itself is not sufficient to guarantee existence of X_i close to X_0, even asymptotically.

The formal conditions on the distance function are very mild, in the sense that virtually any choice will work, asymptotically. However, the choice of distance measure may be important in practice, where n is fixed. For numerical variables, it would make sense to choose a scale invariant measure, possibly adjusting for correlations by using the Mahalanobis distance. It may also be a good idea to adjust the weights of the different components according to their correlation with the target variable (assuming only one value is missing), so that components that carry little information are given lower weights. If the missing value is real valued, one could even build a linear regression model for it, and use the size of the regression weights $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{m-1})^T$ in the distance computation. If we had firm belief in the accuracy of such a linear model, we would use $d(X, X') = |\boldsymbol{\beta}^T(X - X')|$ for our distance function. This d, however, violates our assumptions, as it is not topologically equivalent to Euclidean distance, because $d(X, X')$ can be zero, even for $X \neq X'$. Also, if we had this firm belief, we would use the linear model for imputation directly. A possible compromise would be to use the distance measure $d(X, X') = |\boldsymbol{\beta}^T||(X - X')| = \sum_i |(X_i - X'_i)\beta_i|$, which penalizes each component's deviation by multiplying it with the size of the corresponding regression weight. This d is topologically equivalent to Euclidean distance, provided all β's are nonzero. Similar regression procedures could be defined for ordinal or even nominal data.

Clearly, finding an optimal distance function is a complex problem by itself. From our point of view, a strong point of the RKNN algorithm is its simplicity, together with favourable asymptotic properties, and in practice it may not be worthwhile to build complex regression models in order to optimise the distance measure.

5. Experiment

To shed some light on the issues of convergence speed and distance functions, we present a simulation experiment.

Let $\xi_1, \ldots, \xi_7 \approx N(0, 1)$ be independent standard normal variables, and let $Z_1, \ldots, Z_7$ be a random walk defined by $Z_1 = \xi_1$ and $Z_j = Z_{j-1} + \xi_j$, for $j = 2, \ldots, 7$. For the experiment, we generate n independent data vectors, so that $Z_{i,j}$ is the value of the ith simulated random walk, at step j. We let $Z_{1,7}$ be the only missing value, and investigate the algorithm's ability to impute it, by estimating the variance of $Z_{1,7} - Z_{I,7}$. (The mean of this difference is 0, by construction.) To estimate the variance, we generate 10.000 such Z-matrices, and include all points within the k-neighbourhood in the estimation.

For distance measurements we test Euclidean distance d_E, as well as a regression-based distance d_R. For this random walk Z_6 is a sufficient statistic for Z_7 and d_R is given by $d_R(X, X') = |X_6 - X'_6|$. Table 1 gives the estimated variances for the two distance functions, and different values of n.

The optimal variance value is 2, due to the randomness of the last step in the random walk, and the results appear to confirm convergence toward this value. As we would expect, the optimal regression-based distance measure d_R gives faster convergence.

6. Conclusion

We address the general problem of imputing missing values from a data set of multivariate independent identically distributed variables. We follow the argument of Rubin (1996) that the goal should be to estimate the probability distribution of the missing values, in such a way that the i.i.d. property of the data set is preserved. We prove that the RKNN algorithm with $k = \sqrt{n}$ achieves this asymptotically, under the MAR assumption together with a minimum probability of observations being complete. An experiment confirms the theoretical results, and illustrates how an optimised distance function performs relative to an uninformed one.

References

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer, Berlin.
Hawthorne, G., Elliott, P., 2005. Imputing cross-sectional missing data: comparison of common techniques. Aust. N.Z.J. Psychiatry 39, 583–590.
Little, R.J.A., 1988. Missing-data adjustments in large surveys. J. Bus. Econom. statist. 6 (3),
Reilly, M., 1993. Data-analysis using hot deck multiple imputation. Statistician 42 (3), 307–313.
Rubin, D.B., 1996. Multiple imputation after 18+ years. J. Amer. Statist. Assoc. 91 (434), 473–489.
Rubin, D.B., 2004. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.
Troyanskaya, O., et al., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17 (6), 520–525.
Twisk, J., de Vente, W., 2002. Attrition in longitudinal studies: how to deal with missing data. J. Clinical Epidemiology 55, 329–337.