Computational Statistics and Data Analysis 90 (2015) 84–99

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Improved methods for the imputation of missing data by nearest neighbor methods

Gerhard Tutz [a,*], Shahla Ramzan [b]

[a] Ludwig-Maximilians-Universität München, Akademiestraße 1, D-80799 München, Germany
[b] Ludwig-Maximilians-Universität München, Ludwigstraße 33, D-80539 München, Germany

ARTICLE INFO

Article history:
Received 10 November 2014
Received in revised form 23 April 2015
Accepted 23 April 2015
Available online 30 April 2015

Keywords:
Kernel function
Weighted nearest neighbors
Cross-validation
Weighted imputation
MCAR

ABSTRACT

Missing data raise problems in almost all fields of quantitative research. A useful nonparametric procedure is the nearest neighbor imputation method. Improved versions of this method are presented. First, a weighted nearest neighbor imputation method based on L_q distances is proposed. It is demonstrated that the method tends to have a smaller imputation error than other nearest neighbor estimates. Then weighted nearest neighbor imputation methods that use distances for selected covariates are considered. The careful selection of distances that carry information about the missing values yields an imputation tool that can outperform competing nearest neighbor methods. This approach performs well, especially when the number of predictors is large. The methods are evaluated in simulation studies and with several real data sets from different fields.

1. Introduction

Missing data have always been a challenge to researchers. Since the 1980s, many techniques to impute missing data have been proposed, for example, Little and Rubin (2002) and Schafer (2010). Broadly speaking, the methods for filling in an incomplete data matrix can be divided into two main categories, single imputation and multiple imputation (Little and Rubin, 2002). A well-known and computationally simple method for the imputation of missing data is mean substitution. However, its disadvantage is that the correlation structure among the predictors is ignored. An alternative is the *nearest neighbors* (NNs) approach, which uses observations in the neighborhood to impute missing values. NNs as a nonparametric concept in discrimination dates back to Fix and Hodges (1951). The approach has been successfully used to impute data in gene expression (Troyanskaya et al., 2001; Atkeson et al., 1997; Hastie et al., 1999), machine learning (Batista and Monard, 2002), medicine (Waljee et al., 2013), forestry (Eskelson et al., 2009; Hudak et al., 2008), and compositional data (Hron et al., 2010).

An important aspect in missing data imputation is the pattern of missing values because it determines the selection of an imputation procedure. Little and Rubin (2002) defined three categories of missing data, missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). MCAR refers to data in which the probability of an observation being missing does not depend on the variable itself or on any other variable in the data set (Little and Rubin, 2002; Allison, 2001). Most imputation methods assume the data to be at least MAR, if not MCAR, and so does the NN method. Before considering NN approaches in detail, a few general remarks on the method are warranted. The main

* Corresponding author. Tel.: +49 89 2180 3044; fax: +49 89 2180 5308.
E-mail addresses: tutz@stat.uni-muenchen.de (G. Tutz), shahla.ramzan@stat.uni-muenchen.de (S. Ramzan).

http://dx.doi.org/10.1016/j.csda.2015.04.009
0167-9473/

approach for imputation is multiple imputation because it yields valid inferences under several basic conditions. In contrast to NN approaches, multiple imputation provides consistent estimates under MAR conditions. However, NN approaches are useful in high dimensional problems in which multiple imputation cannot be applied and complete cases do not generate enough data.

Several NN approaches to imputation have been considered in the literature. Hastie et al. (1999) proposed using weights based on the Euclidean distance for NN selection. Troyanskaya et al. (2001) compared the k nearest neighbor imputation (KNNimpute) with the mean imputation and singular-value decomposition (SVD) techniques for gene expression data. Their simulation studies showed that the method performs well compared to mean imputation and SVD approaches, as also demonstrated by Troyanskaya et al. (2003). In a comparative study of single imputation methods, Malarvizhi and Thanamani (2012) found that median or standard deviation substitution perform better than mean substitution. A further comparison of KNNimpute with mean, ordinary least squares (OLS) and partial least squares (PLS) imputation methods in microarray data was that of Nguyen et al. (2004). In that study, the good performance of KNNimpute was demonstrated.

Several alternative procedures have been proposed that rely on basic concepts to impute values by building averages over qualifying neighbors, as described in Ouyang et al. (2004), Kim et al. (2004), Sehgal et al. (2005), and Scheel et al. (2005). Liew et al. (2011) and Moorthy et al. (2014) reviewed the available methods and algorithms, with a focus on gene expression data. Johansson and Hakkinen (2006) proposed *WeNNI*, which utilizes continuous weights in the NN imputation procedure. Bøet al. (2004) and Wasito and Mirkin (2005) proposed a NN procedure with a modification based on the least squares method. Other variants include the local least squares (LLSimpute) method of Kim et al. (2005), the sequential local least squares (SLLSimpute) method of Zhang et al. (2008), and the iterative LLS (ILLSimpute) of Cai et al. (2006).

A drawback of NN methods is that their performance depends on k. For example, KNNimpute typically performs well when k is between 5 and 10, but the performance deteriorates for larger values of k (Yoon et al., 2007). Here we propose a localized approach to missing data imputation that uses a weighted average of NNs based on L_q distances. For the high-dimensional case, we propose a new distance that explicitly uses the correlation among variables. The method automatically selects the relevant variables that contribute to the distance and thus does not depend on k.

The paper is organized as follows: In Section 2, the L_q distance is used to define a weighted imputation estimate. In a simulation study, the weighted approach is compared to the un-weighted approach. In Section 3, the weighted imputation with the selection of predictors is introduced and compared to alternative imputation techniques. In Section 4, several applications using real data sets are demonstrated.

2. Weighted neighbors

When using NNs to impute data, two preliminary steps must be carried out. First, one has to define the NNs, that is, how they are computed and which distance measures are to be used. Second, one has to determine how these NNs are used to obtain an imputed value. The choices to be made in these two steps are considered in the following sections.

2.1. Distances and computation of nearest neighbors

Let n observations on p covariates be collected. The corresponding $n \times p$ data matrix is given by $\mathbf{X} = (x_{is})$, where x_{is} denotes the ith observation of the sth variable. Let $\boldsymbol{O} = (o_{is})$ denote the corresponding $n \times p$ matrix of dummies with entries

$$o_{is} = \begin{cases} 1 & \text{if } x_{is} \text{ was observed} \\ 0 & \text{for missing value.} \end{cases}$$

Distances between two observations $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, represented by rows in the data matrix, can be computed by using the L_q metric for the observed data, given by

$$d_q(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left[\frac{1}{m_{ij}} \sum_{s=1}^{p} |x_{is} - x_{js}|^q I(o_{is} = 1) I(o_{js} = 1) \right]^{1/q}, \tag{1}$$

where $m_{ij} = \sum_{s=1}^{p} I(o_{is} = 1) I(o_{js} = 1)$ denotes the number of valid components in the computation of distances. The indicator function $I(a)$, which is used in the definition, has the value 1, if a is true and 0 otherwise. The computation of distances does not use all the components of the vectors but only those for which observations in both vectors are available. The components included in the computation of neighbors are given by $C_{ij} = \{s : I(o_{is} = 1) I(o_{js} = 1) = 1\}$. The distances define the NNs when imputing a specific value. It should be noted that the number of components varies over the sample because C_{ij} depends on the specific observation for which imputations are to be made.

Similar concepts to define distances and therefore NNs were described by Hastie et al. (1999), Troyanskaya et al. (2001), Myrtveit et al. (2001), and Kim et al. (2005). In those studies the Euclidean distance was used. Alternatives to compute distances in gene expression studies are based on the Pearson correlation (Dudoit et al., 2002; Bøet al., 2004) and on covariance estimates (Sehgal et al., 2005).