Chemometrics and Intelligent Laboratory Systems 120 (2013) 106–115

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Comparison of five iterative imputation methods for multivariate classification

Yushan Liu, Steven D. Brown *

Department of Chemistry and Biochemistry, University of Delaware, Brown Laboratory, 163 The Green, Newark, DE 19716, USA

ARTICLE INFO

Article history:
Received 18 September 2012
Received in revised form 7 November 2012
Accepted 11 November 2012
Available online 22 November 2012

Keywords:
Multivariate imputation
Iterative imputation
Covariance criterion
Classification criterion

ABSTRACT

Imputation methods are often used to fill the missing values in an incomplete data set before applying multivariate statistical methods. In this paper, five iterative imputation methods are compared. These include general iterative principal component imputation (GIP), singular value decomposition imputation (SVD), regularized expectation maximization with multiple ridge regression (r-EM), regularized expectation maximization with truncated total least squares (t-EM), and multiple imputation by chained equations (MICE). Two evaluation criteria (covariance change and classification error change) are determined to evaluate imputation performance on one simulated dataset and two published datasets. No single imputation method emerged as the overall best in all cases examined. The r-EM imputation method performs well when the missing proportion is under 20%, judging from results obtained from both real datasets examined. If the percentage of the missing data is above 20%, however, the purpose behind analysis of a dataset should be considered carefully before choosing an imputation method.

1. Introduction

Missing data remain a significant challenge for successfully applying any state-of-the-art multivariate analysis technique because most, if not all, multivariate methods were developed under the assumption that the data to be used as input are complete. Imputation is often used to estimate these missing elements [1,2]. The imputed dataset is then ready for use with a variety of multivariate methods.

In the field of chemometrics, different imputation methods have been widely applied to incomplete datasets. Most methods focus on estimating a robust mean and covariance [1–3], on building PCA models [4–9] and on constructing regression models [5,10–14]. Comparatively, imputation prior to multivariate classification seems to have received less attention, because classification performance is considered less sensitive to the choice of imputation methods when the dataset does not contain outlying values [15,16]. In most classification studies of incomplete datasets, a lack of investigation of the accuracy of imputed data is common. The performance of an imputation method is generally based solely on classification accuracy [17,18] and imputation accuracy is neglected. However, classification accuracy and imputation accuracy are both important because there is no guarantee that more accurately imputed data will result in better classification performance and vice versa [16].

To investigate the effects of different imputation methods on multivariate classification analysis adequately, two criteria are used in this study. One criterion measures imputation accuracy and the other measures classification accuracy. RMSE (root-mean-square error) measures are widely used in evaluating imputation accuracy [1,19], but their results are not consistent with the results of classification accuracy, which is usually represented by the classification error. In this study, the root-mean-square deviation of the estimated covariance matrix of the imputed data matrix from the population covariance matrix [14] is used to assess imputation accuracy instead of evaluating the imputed values by RMSE measures, as it is related closely to the discriminant classifier used in this study and because it evaluates the changes in covariance directly. Its correlation with classification accuracy has not been previously discussed in the literature.

To explore the relation between the two types of accuracy of imputation, the missing elements in several incomplete, multivariate datasets are imputed using five common iterative imputation methods, including general iterative principal component imputation (GIP) [14], singular value decomposition imputation (SVD) [14], regularized expectation maximization with multiple ridge regression method (r-EM) [3], regularized expectation maximization with truncated total least squares (t-EM) [3] and multiple imputation by chained equations (MICE) [20]. Their performance is assessed via the change in the covariance matrix and the change in classification error caused by the imputation, and the relation of these two criteria is considered. Balancing strategies as well as general suggestions in choosing imputation methods for practical incomplete datasets are discussed.

2. Iterative imputation methods

Missing mechanisms, which describe how the missing variables are related to the underlying values of the variables in the dataset,

* Corresponding author. Tel.: +1 302 831 6861; fax: +1 302 831 6335.
E-mail address: sdb@udel.edu (S.D. Brown).

0169-7439/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2012.11.010

are usually considered before applying any imputation methods. These mechanisms are categorized as follows:

1. Missing at random (MAR), when the distribution of missing data for a variable only depends on observed data, but does not depend on the missing data itself. This is the most common assumption when one encounters missing data because most efficient imputation methods are based on it. In practice, the MAR pattern occurs commonly in the fields in which data are from large surveys, such as those in social science or economics [2], but it occurs far less often in chemistry.
2. Missing completely at random (MCAR), when the distribution of missing data for a variable does not depend on observed or on any missing data. This mechanism is fairly common in chemical data. For example, a significant level of iron is being sought in a decomposition test of a chemical compound. Missing data for iron can be considered as MCAR if certain levels of iron are missing entirely and if no other elements present in the compound are correlated with iron at those levels [21].
3. Not missing at random (NMAR), when the distribution of missing data for a variable depends on both the observed as well as the missing data. The NMAR mechanism is also common in chemistry, particularly when some data values are below the detection limit.

Previous studies have shown that imputation methods are specific to a particular missing mechanism and cannot be simply applied to multivariate data where the missing elements show a different mechanism [1]. It is therefore essential to distinguish the probable missing mechanism when missing data arises.

To focus on the classification performance of different imputation methods, the missing mechanism of all datasets is restricted to MCAR in this study because this mechanism is common in chemistry and because most state-of-art imputation methods can be applied directly or with minor modifications. All of these imputation methods are based on an iterative algorithm:

1. An initial guess for missing data is provided. Any guess is possible, but in practice, mean values of each variable from the available data are preferable [14]. The missing elements are filled with these initial guesses and a complete dataset is created.
2. Model parameters are estimated for the complete dataset generated in Step 1. For different imputation methods, the model parameters to be estimated vary. Details of the choice of parameters are discussed below.
3. The estimated model parameters are used to find the conditional expectation of the missing elements. The conditional expectation is calculated from available data and those estimated parameters.
4. The missing elements are replaced with their expectations obtained from Step 3.

Steps 2 through 4 are iterated until consecutive iterates of imputed values are within a specified tolerance. In this work, a tolerance of 10^{-6} was used. The five imputation methods used here follow these steps, but these estimate different model parameters and calculate different conditional expectations. Among the five imputation methods investigated here, MICE, r-EM and t-EM require missing data to be MCAR, but GIP and SVD are not tied to a specific mechanism for the missing entries. The reasons are discussed below.

The first method considered in this paper, the GIP imputation algorithm, introduces iterative steps in Dear's principal component (DPC) imputation method [22]. In DPC imputation, the first principal component is estimated from the covariance matrix of all available data and the missing elements are replaced by the nearest point on the first principal component, without any need for iteration [23]. Suppose that x_{ij} is an element from the $m\times p$ dimensional data matrix $\mathbf{X}$. To treat all variables equally, $\mathbf{X}$ is standardized to $\mathbf{Z}$, where each element $z_{ij}=(x_{ij}-\bar{x}_j)/\sqrt{s_j^2}$ and where $\bar{x}_j$ and $\sqrt{s_j^2}$ are the mean and standard deviation of the available data for the jth variable. A missing indicator matrix, $\mathbf{R}=\{r_{ij}\}$, is then defined as

$$r_{ij}=\begin{cases}1 & \text{if } x_{ij} \text{ is observed}\\ 0 & \text{if } x_{ij} \text{ is missing}\end{cases} \quad (1)$$

The next step is to construct the correlation matrix $\mathbf{C}$ by using the available data only and to obtain the largest eigenvalue of $\mathbf{C}$, namely $\lambda_1=\max_j(\lambda_j)$, and its associated eigenvector η_{1j}. The first principal component score for the ith sample is

$$\gamma_i=\sum_{j=1}^{p}\eta_{1j}z_{ij}r_{ij} \quad (2)$$

and the missing elements are replaced by the points that are closest to the ith sample; that is,

$$\hat{z}_{ij}=\begin{cases}z_{ij} & \text{if } r_{ij}=1\\ \eta_{1j}\gamma_i & \text{if } r_{ij}=0\end{cases} \quad (3)$$

Eqs. (2) and (3) are repeated for all samples having missing elements, and then $\hat{\mathbf{Z}}$ is rescaled to $\hat{\mathbf{X}}$, where $\hat{\mathbf{X}}$ is the complete dataset with estimated values for missing data. DPC imputation requires no distributional assumptions for its use, but it may provide poor estimates of the correlation matrix because only the available data are used. GIP imputation can overcome this shortcoming through introducing initial guesses for missing data. Here, the estimated correlation matrix $\mathbf{C}$ is calculated from the full data matrix rather than only the available data. Eqs. (2) and (3) are then iterated until consecutive imputed values are within the specified tolerance.

The second imputation method, one proposed by Krzanowski [24], uses the well-known singular value decomposition (SVD) to impute the missing data. The SVD imputation algorithm is easy to implement because singular value decomposition is the only algorithm involved [13]. Similar to GIP imputation, SVD imputation does not require any multivariate distributional assumption for its use. Again suppose that x_{ij} is a missing element from the $m\times p$ dimensional data matrix $\mathbf{X}$. The ith row containing x_{ij} is deleted from $\mathbf{X}$ and the SVD is calculated on the remaining $(m-1)\times p$ matrix $\mathbf{X}^{-i}$, so that

$$\mathbf{X}^{-i}=\bar{\mathbf{U}}\bar{\mathbf{D}}\bar{\mathbf{V}}' \quad (4)$$

where $\bar{\mathbf{U}}$ and $\bar{\mathbf{V}}$ are orthogonal matrices of dimension $(m-1)\times(m-1)$ and $p\times p$, respectively, and $\bar{\mathbf{D}}=\mathrm{diag}\{\mathbf{d}_1,\ldots,\mathbf{d}_p\}$.

Similarly, the jth variable is deleted from $\mathbf{X}$ and the SVD of the remaining $m\times(p-1)$ matrix $\mathbf{X}_{-j}$ is obtained, so that

$$\mathbf{X}_{-j}=\tilde{\mathbf{U}}\tilde{\mathbf{D}}\tilde{\mathbf{V}}' \quad (5)$$

where $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}$ are orthogonal matrices of dimension $m\times m$ and $(p-1)\times(p-1)$, respectively, and $\tilde{\mathbf{D}}=\mathrm{diag}\{\tilde{\mathbf{d}}_1,\ldots,\tilde{\mathbf{d}}_{p-1}\}$.

Then the value of the missing element x_{ij} is calculated by

$$\hat{x}_{ij}=\sum_{t=1}^{p-1}\left[\tilde{u}_{it}\tilde{d}_t^{1/2}\right]\left[\bar{v}_{jt}\bar{d}_t^{1/2}\right] \quad (6)$$

where $\tilde{u}$ and $\bar{v}$ are elements in matrices $\tilde{\mathbf{U}}$ and $\bar{\mathbf{V}}$, respectively.

In this way, x_{ij} is estimated using the maximum information available in the data, avoiding bias [25]. If $\mathbf{X}$ contains more than one missing element, an initial guess is provided for the missing data other than the entry x_{ij}. Eqs. (4) to (6) are then iterated until consecutive iterates of each of the imputed values are within the specified tolerance.

In MICE imputation, initial guesses for all missing elements x_{ij} are provided for the $m\times p$ incomplete dataset $\mathbf{X}$. For each variable with missing elements, $\mathbf{x}_j$, the data are split into two sub-vectors: $\mathbf{x}_{ja}$ a sub-vector that contains all available data, and $\mathbf{x}_{jm}$ a sub-vector that