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

contains all missing data. The available sub-vector $\mathbf{x}_{ja}$ is regressed on all other variables, which are restricted to the samples in $\mathbf{x}_{ja}$; that is

$$\mathbf{x}_{ja} = f\left(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{j-1}, \mathbf{x}_j, \cdots, \mathbf{x}_p\right) \quad 7$$

The missing sub-vector $\mathbf{x}_{jm}$ is then predicted from the regression and its missing entries are replaced with the predictions from the regression. The regression procedure is repeated for all variables with missing elements. After all missing elements are imputed, the regressions and predictions are repeated until consecutive iterates are within the specified tolerance for each of the imputed values. The procedure is illustrated in Fig. 1.

Because a series of regression models are run in which each variable associated with the missing values is modeled conditionally on the other variables in the data, MICE can deal with different variable types [26]. While the regression model f in Eq. (7) can be selected from among linear, logistic and nonlinear models, a linear regression model is used to perform that modeling in this study because the linear model is the most common choice for imputing normally distributed variables [26]. Discussions of the other advantages of MICE over other imputation methods can be found in the literature [20,27].

The remaining two imputation algorithms used in this study perform imputation based on versions of regularized expectation maximization (EM) [3]. In the general EM algorithm, an estimation of the parameters of incomplete datasets is obtained from maximizing the likelihood on available data. For instance, suppose that the probability distribution of incomplete, multivariate normally distributed data $\mathbf{X}$ is under investigation. The EM algorithm starts with initial estimates of the mean and the covariance matrix of $\mathbf{X}$ to estimate the distribution. As in MICE imputation, the regression coefficients $\mathbf{B}$ of the variables with missing elements on the variables with available values are computed with the estimated mean and covariance matrix. The missing data are then imputed with their conditional expectation values, which are the product of the available values and the estimated regression coefficients, and $\mathbf{X}$ is updated with imputed values. This is the expectation step (E-step). The mean and covariance matrix are estimated again based on the updated value for $\mathbf{X}$, which is called the maximization step (M-step). The E-step and M-step are iterated until the imputed values as well as their estimated means and covariance change within a specified tolerance [28]. The estimated mean and covariance describe the multivariate probability distribution of this incomplete dataset.

Steps described above indicate how the general EM algorithm can be used to impute incomplete datasets. However, the inverse of the estimated covariance that is used to obtain the regression coefficients in this algorithm can be ill-defined, which can cause the estimation of the regression coefficients $\mathbf{B}$ to fail. In the first regularized EM method used in this study, the conditional maximum likelihood estimate of the regression parameters used for calculating conditional expectation values can be estimated as

$$\mathbf{B} = \hat{\boldsymbol{\Sigma}}_{aa}^{-1}\hat{\boldsymbol{\Sigma}}_{am} \quad 8$$

where $\mathbf{B}$ is the matrix of regression coefficients, $\hat{\boldsymbol{\Sigma}}_{am}$ consists of the estimated cross-covariances of variables for which the values are available with those variables for which the values are missing, and the $\hat{\boldsymbol{\Sigma}}_{aa}^{-1}$ term in Eq. (8) is calculated with a regularized inverse

$$\hat{\boldsymbol{\Sigma}}_{aa}^{-1} = \left(\hat{\boldsymbol{\Sigma}}_{aa} + h^2 diag\left(\hat{\boldsymbol{\Sigma}}_{aa}\right)\right)^{-1} \quad 9$$

where h is a regularization parameter. This regularized EM method, in which the inverse of a matrix ($\hat{\boldsymbol{\Sigma}}_{aa}$) is replaced with the inverse of the sum of the matrix and a multiple of a positive definite matrix (diag($\hat{\boldsymbol{\Sigma}}_{aa}$)), is called ridge EM (r-EM). In this way, direct calculation using the ill-defined inverse $\hat{\boldsymbol{\Sigma}}_{aa}^{-1}$ can be avoided. At each step, the missing values are estimated by $\hat{\boldsymbol{x}}_m$ where

$$\hat{\mathbf{x}}_m = \hat{\boldsymbol{\mu}}_m + (\mathbf{x}_a - \hat{\boldsymbol{\mu}}_a)\mathbf{B} \quad 10$$

where $\hat{\boldsymbol{\mu}}_m$ is the estimated mean vector for which values are missing, $\hat{\mu}_a$ is the estimated mean vector for which values are available, and where $\mathbf{x}_a$ is the matrix containing the available values of the matrix $\mathbf{X}$. Eqs. (8) to (10) are repeated for all missing values until consecutive iterates produce imputed data whose changes in values fall within the specified tolerance.

Another regularized EM method (t-EM) uses a truncated total least squares algorithm, in which the regression coefficients $\mathbf{B}$ in Eq. (9) are computed from the first q principal components of the overall covariance matrix instead of the scaled $\hat{\boldsymbol{\Sigma}}_{aa}$ [3].

$$\mathbf{B} = \left(\mathbf{V}_{aq}^{+}\right)^{\mathrm{T}}\left(\mathbf{V}_{mq}^{\mathrm{T}}\right) \quad 11$$

Here, $\mathbf{V}_{aq}$ and $\mathbf{V}_{mq}$ are sub-matrices of the matrix $\mathbf{V}$ that contains as its columns the orthogonal eigenvectors of the estimated covariance matrix $\hat{\boldsymbol{\Sigma}}$; that is, $\hat{\boldsymbol{\Sigma}} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}$. The sub-matrix $\mathbf{V}_{aq}$ contains the first q eigenvectors of V that belong to the variables for which the values are available, where $\mathbf{V}_{\mathrm{aq}}^{+} = (\mathbf{V}_{\mathrm{aq}}^{\mathrm{T}}\mathbf{V}_{\mathrm{aq}})^{-1}\mathbf{V}_{\mathrm{aq}}^{\mathrm{T}}$ is the pseudoinverse of the matrix of $\mathbf{V}_{aq}$; the sub-matrix V_{mq} contains the first q eigenvectors of V that belong to the variables for which the values are missing. This algorithm is called truncated total least squares because the regression

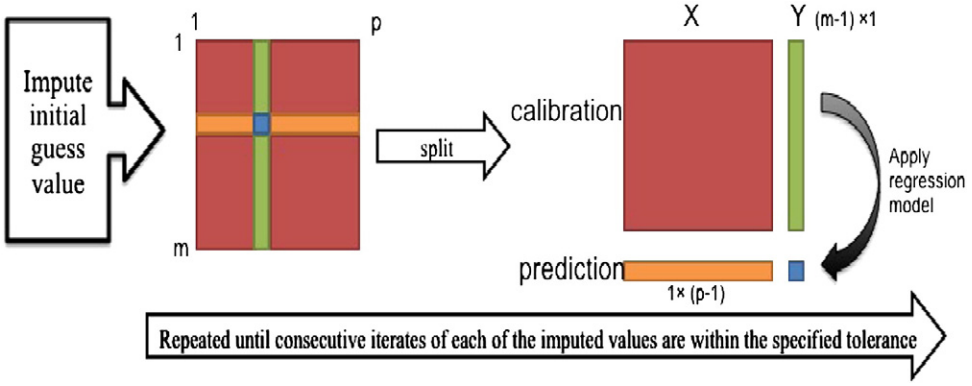

Fig. 1. Iterative steps for the MICE imputation method. The number of samples is m and the number of variables is n. The small (shown in blue in the web version) square in the middle represents a missing entity; the horizontal rectangular stripe (shown in orange) represents a sample with a missing entity, the vertical (shown in green) rectangle represents the variable with a missing entity, and the large (shown in red) rectangle represents data with no missing entities.

coefficients **B** are calculated from only the first q eigenvectors of the estimated overall covariance matrix instead of the full covariance matrix itself. This method is an alternative approach for regularization of ill-conditioned covariance matrices [29].

3. Imputation criteria

Two criteria are utilized to compare the performance of different imputation methods. One criterion, called the covariance criterion (ψ_{cov}), is a statistical quantity that measures the accuracy of the estimated population covariance by comparing the estimate to the population covariance matrix. The other criterion, called the classification criterion (ψ_{err}), measures changes in classification error of the complete dataset before the missing values are introduced (the complete dataset) and the complete dataset after imputation procedure (the imputed dataset).

3.1. Covariance criterion

When the original true values are known, several statistical measures can be used to measure the similarity of imputed values and original true values, such as RMSE and normalized RMSE [1,19]. These statistical quantity measures are widely applied to examine the performance of imputation methods because they are intuitive metrics to evaluate the deviation of estimated values for missing elements. The covariance criterion used in this study, which is similar to the RMSE metric, can be viewed as a root-mean-square deviation of the estimated covariance matrix of the imputed data matrix from the covariance matrix of the complete dataset before the missing values are introduced [14]. The metric ψ_{cov} is calculated as

$$\psi_{cov} = \mathrm{tr}\left\{\left(\boldsymbol{\Sigma}_{\mu} - \hat{\boldsymbol{\Sigma}}_{\mu t}\right)\left(\boldsymbol{\Sigma}_{\mu} - \hat{\boldsymbol{\Sigma}}_{\mu t}\right)\right\}/p^2 \quad (12)$$

where 'tr' is the trace of a matrix, $\boldsymbol{\Sigma}_{\mu}$ represents the covariance matrix of the complete dataset before the missing values are introduced, $\hat{\boldsymbol{\Sigma}}_{\mu t}$ represents the covariance matrix of the complete dataset after imputation (the imputed dataset), and where p is the number of variables in the dataset. Generally speaking, the higher the value of ψ_{cov}, the larger the difference between covariance matrices of the imputed and complete datasets. There is no established statistical threshold to provide a measure of the quality of the imputation, but the metric can be useful for direct comparisons, as was done in [14].

3.2. Classification criterion

The goal of many studies involving missing data is classification, and to perform a classification, it is necessary to determine the classes of samples from variables [30]. Therefore, the impact of imputation of any missing data on classification of a dataset is also examined here. In this paper, misclassification error is calculated for a Bayes quadratic discriminant classifier (QDA). This classifier fits a multivariate normal density distribution to each class with covariance estimates stratified by class. The quadratic discriminant function is defined as

$$\mathbf{d}_k(x) = \ln\left(\pi_k \cdot f(x|k)\right) = \ln(\pi_k) - \frac{1}{2}\ln(\det(\boldsymbol{\Sigma}_k)) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k(\mathbf{x}-\boldsymbol{\mu}_k) \quad (13)$$

where π_k are the prior probabilities for class k, $\boldsymbol{\Sigma}_k$ is the covariance matrix for class k, and where μ_k is the mean vector for class k. Detailed discussion of this classifier can be found elsewhere [30,31]. Unlike the linear discriminant (LDA) classifier, the QDA classifier does not require that the covariance matrices for all classes are similar. Because the covariance of each class is typically unknown for incomplete data, it is reasonable to apply a classifier based on QDA instead of LDA.

For QDA a training data set was selected and the derived classification rule was used to predict the class membership of the remaining data (test data). The training (2/3 of the samples) and test sets were selected randomly with stratification; that is, samples were selected randomly such that both training and test sets had roughly the same class membership proportions as those in the full data set. The classification error was then computed for all classes jointly. Since the error will depend on the choice of the test data, the procedure was repeated 100 times. The mean misclassification error of those 100 runs is reported as the final total classification error. The classification criterion (ψ_{err}) is then defined as

$$\psi_{err} = (e_2 - e_1)/e_1 \quad (14)$$

where e_1 and e_2 represent the classification error of the complete dataset and the imputed dataset respectively. As with ψ_{cov}, there is no established statistical threshold for ψ_{err}. The lower the value of ψ_{err}, the less the difference between classification errors of the complete dataset before the missing values are introduced (the complete dataset) and those of the complete dataset after imputation procedure (the imputed dataset).

3.3. Relation between the two criteria

The covariance criterion and classification criterion provide different assessments of imputation performance. To investigate the relation between the two criteria, the covariance structure of each class in dataset is examined first. The covariance structure is used in QDA to calculate the discriminant boundary. Based on [32–34], the eigen decomposition of the covariance matrix for the kth class can be expressed as

$$\boldsymbol{\Sigma}_k = \mathbf{D}_k\boldsymbol{\Lambda}_k\mathbf{D}_k^T = \mathbf{D}_k\lambda_k\mathbf{A}_k\mathbf{D}_k^T = \lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T \quad (15)$$

where $\mathbf{D}_k$ is the matrix of eigenvectors and where $\boldsymbol{\Lambda}_k$ is a diagonal matrix with the eigenvalues of $\boldsymbol{\Sigma}_k$ on the diagonal. The orientation of the covariance matrix for the kth class is determined by $\mathbf{D}_k$ and the volume of the class covariance matrix for the kth class is represented by the determinant of Λ_k, where Λ_k can be further divided into $\boldsymbol{\lambda}_k$ and A_k,

$$\boldsymbol{\Lambda}_k = \lambda_k\mathbf{A}_k = |\boldsymbol{\Sigma}_k|^{1/d}\mathbf{A}_k \quad (16)$$

Here, $\mathbf{A}_k$ is a diagonal matrix such that $|\mathbf{A}_k|$, the determinant of A_k, is 1, and with normalized eigenvalues of $\boldsymbol{\Sigma}_k$ on the diagonal in decreasing order; that is, $\mathbf{A}_k = \mathrm{diag}\{\alpha_{1k}, \ldots, \alpha_{pk}\}$ such that $\alpha_{1k} \geq \alpha_{2k} \geq \ldots \geq \alpha_{pk} > 0$.

Because the eigenvalues α are normalized, $\boldsymbol{\lambda}_k$ alone represents the volume of the covariance matrix for the kth class. Furthermore, the ratio of the diagonal elements of $\mathbf{A}_k$ is related to the shape of the covariance matrix. A ratio equal to 1 indicates that the covariance is circular, with equal variable contributions to covariance, while larger ratios indicate a dominant direction, and in the extreme case, a ratio approaching infinity means that the shape of the covariance matrix collapses to a line. The parameters in Eq. (15) together describe the covariance structure of the kth class; that is, $\mathbf{D}_k$ determines the orientation of the kth class, $\boldsymbol{\lambda}_k$ determines the volume of the covariance matrix of the kth class and $\mathbf{A}_k$ determines its shape.

A classic variable selection method, Wilks' lambda, was used to reveal the relation between two criteria. Wilks' lambda, often used in a stepwise analysis to identify multiple variables that maximize separation of classes from one another [35,36], is defined as the ratio of the determinant of the within-class covariance matrix over the determinant of the total covariance matrix. Based on the definition, low values of Wilks' lambda indicate variables that better discriminate classes. Because Wilks' lambda is calculated from the covariance matrix and can determine the most significant variables of a dataset for

classification, it connects the two criteria used in this study. The process of stepwise multivariate Wilks' lambda analysis is as follows:

1. Calculate the Wilks' lambda for each variable in a dataset $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p]$.
2. The variable that has the best discriminant power (smallest Wilks' lambda in Step 1) is chosen ($\mathbf{x}_t (1 \le t \le p)$) and other variables are combined with x_t separately to form sets of variables ($[\mathbf{x}_t, \mathbf{x}_1]$, $[\mathbf{x}_t, \mathbf{x}_2]$, …, $[\mathbf{x}_t, \mathbf{x}_{t-1}]$, $[\mathbf{x}_t, \mathbf{x}_{t+1}]$, …, $[\mathbf{x}_t, \mathbf{x}_p]$).
3. Calculate multivariate Wilks' lambda for each set of variables in Step 2. The set has the best discriminant power if its Wilks' lambda is smallest.

The variables from the set that has the best discriminant power are then combined with this set separately to form new sets of variables. Steps 2 and 3 are repeated until all variables are added into the set of variables $\mathbf{X}_{new} = [\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_t, \ldots]$, where $1 \le i, j, t, \ldots \le p$. The variables in $\mathbf{X}_{new}$ are ordered for the classification, from most discriminative to the least.

4. Results and discussion

One simulated and two well-known, real data sets are used in this study. Both real data sets, the Iris dataset and Wine dataset, were obtained from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml). All of the datasets used here have no missing elements, so data elements are first deleted to simulate sets with missing data under the MCAR mechanism, then imputed by using each of the five imputation methods described in Section 2. After the imputed dataset is obtained, both imputation criteria described in Section 3 are calculated for all data sets.

4.1. Simulated dataset

A simulated dataset **X** is generated to reveal the relation between covariance and classification criteria. The dataset contains two variables and consists of matrices defining two classes, **G1** and **G2**, with each class having 50 samples. The variable mean vector of **G1** is [1,0] while the mean for **G2** is [−1,0]. Thus, the mean vector of the simulated dataset **X** is [0,0]. Covariance matrices Σ_1 and Σ_2 are made equal in the two classes [37]. Details of the dataset are shown in Fig. 2.

This simulated dataset is then analyzed. 10% of the missing elements are set in the first variable randomly to generate missing dataset M1 first, and in the second variable to generate set M2. The structure of the missing values in both sets follows the MCAR mechanism even though the missing elements occur only in the first variable in M1 and only in the second variable in M2. This can be achieved if the distribution of missing data in one variable does not depend either on themselves or the observed data, even though the missing elements do not spread randomly over the whole dataset **X**. For both incomplete, simulated datasets M1 and M2, the five iterative imputation methods are used to fill in the missing values. Their performance is evaluated by using the two criteria discussed in Section 3. Comparing the two criteria within M1 or M2 allows examination of how changes in the data covariance matrix arising as a consequence of imputation affect the QDA classification. By comparing covariance and classification criterion results between M1 and M2, changes in covariance and in classification error can be seen when the missing elements appear in different variables, even though the missing mechanism is MCAR for both. The results are shown in Fig. 3.

Comparing the classification criterion ψ_{err} and the covariance criterion ψ_{cov} over sets **M1** (Fig. 3a and b) and **M2** (Fig. 3c and d), it is clear that the two criteria do not always agree. From Fig. 3a and b, the plot of ψ_{err} indicates that SVD imputation is the best method overall (Fig. 3a), but SVD imputation does not show the smallest covariance change ψ_{cov} (Fig. 3b). In contrast, the r-EM method has the smallest ψ_{cov} value after imputation of the **M1** dataset (Fig. 3b), but r-EM imputation has a relatively high ψ_{err} (Fig. 3a), which indicates that classification error increases more than the other imputation methods used on dataset **M1**. In addition to the r-EM and SVD imputation methods in **M1**, results from the other imputation methods also disagree in these two criteria except for t-EM imputation. The t-EM imputation has the highest ψ_{err} and ψ_{cov} metrics. This disagreement between the metrics is even clearer in **M2**, as seen by comparing Fig. 3c and d. Imputation methods producing higher ψ_{err}, such as MICE and r-EM, have lower ψ_{cov} values.

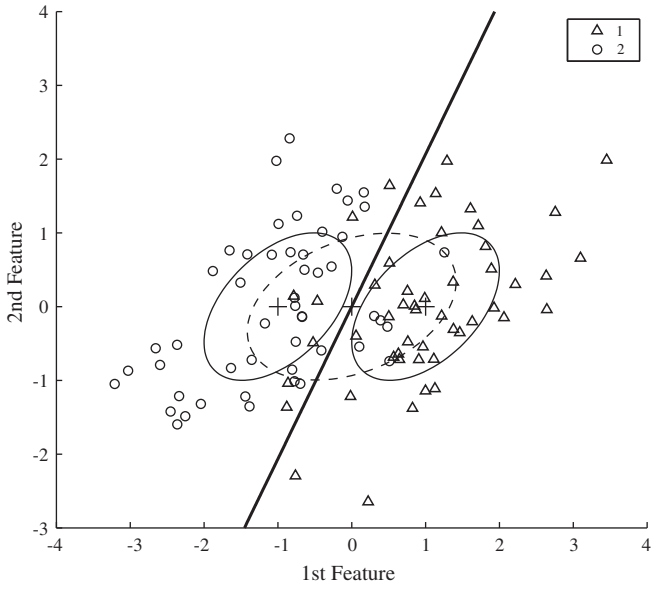

Fig. 2. Covariance structure of the simulated dataset with classes **G1** and **G2**. A circle represents one sample in **G2** and a triangle represents one sample in **G1**. Crosses represent the location of the means. From left to right, the mean centers of **G2**, **X** and **G1** are given, respectively. The two ellipses with solid lines represent covariance matrices of **G1** and **G2**, respectively. The ellipse with a dashed line represents the covariance of **X**. The line represents the discriminant boundary from the Bayes quadratic discriminant classifier (QDA).

If the two criteria are compared between **M1** and **M2** for the same imputation method, GIP imputation for instance, the classification criterion metric for **M1** (Fig. 3a) is much larger than that for **M2** (Fig. 3c), but the covariance criterion metrics in **M1** (Fig. 3b) and **M2** (Fig. 3d) for GIP imputation are comparable. A similar trend can also be found for other imputation methods such as MICE and r-EM. As for SVD and t-EM imputation, even though the values of ψ_{cov} for each method in **M1** (Fig. 3b) and **M2** (Fig. 3d) are notably different compared with values of ψ_{cov} of the other three methods, the metrics still have the same order of magnitude. However, the classification criterion metric for SVD in **M1** increases by nearly two orders of magnitude in **M2**. For t-EM imputation, the classification criterion metric in **M2** is negative, which indicates that the classification error after imputation is *lower* than that obtained using the original (complete) dataset **X**.

The difference between the two criteria is expected because they evaluate different aspects of a dataset, but it is not simple to infer the reason that classification criteria for the same dataset show a large difference when missing elements appear at different variables. To find out the reason for the differences in these two criteria for most of the imputation methods, the covariance matrix was decomposed based on Eq. (15). Table 1 shows all covariance structure parameters for the complete data matrix **X** before removing 10% elements as well as for the two classes contained in **X**: **G1** and **G2**.

Results in Table 1 are consistent with what is seen in Fig. 2. The shape (A_k), the volume (λ_k) and the orientation (D_k) of the two classes contained in **X** are exactly the same. Because the value of λ_k for each class is smaller than that for the whole dataset **X** in Table 1, the volume of each class is smaller than that of the whole dataset, again as

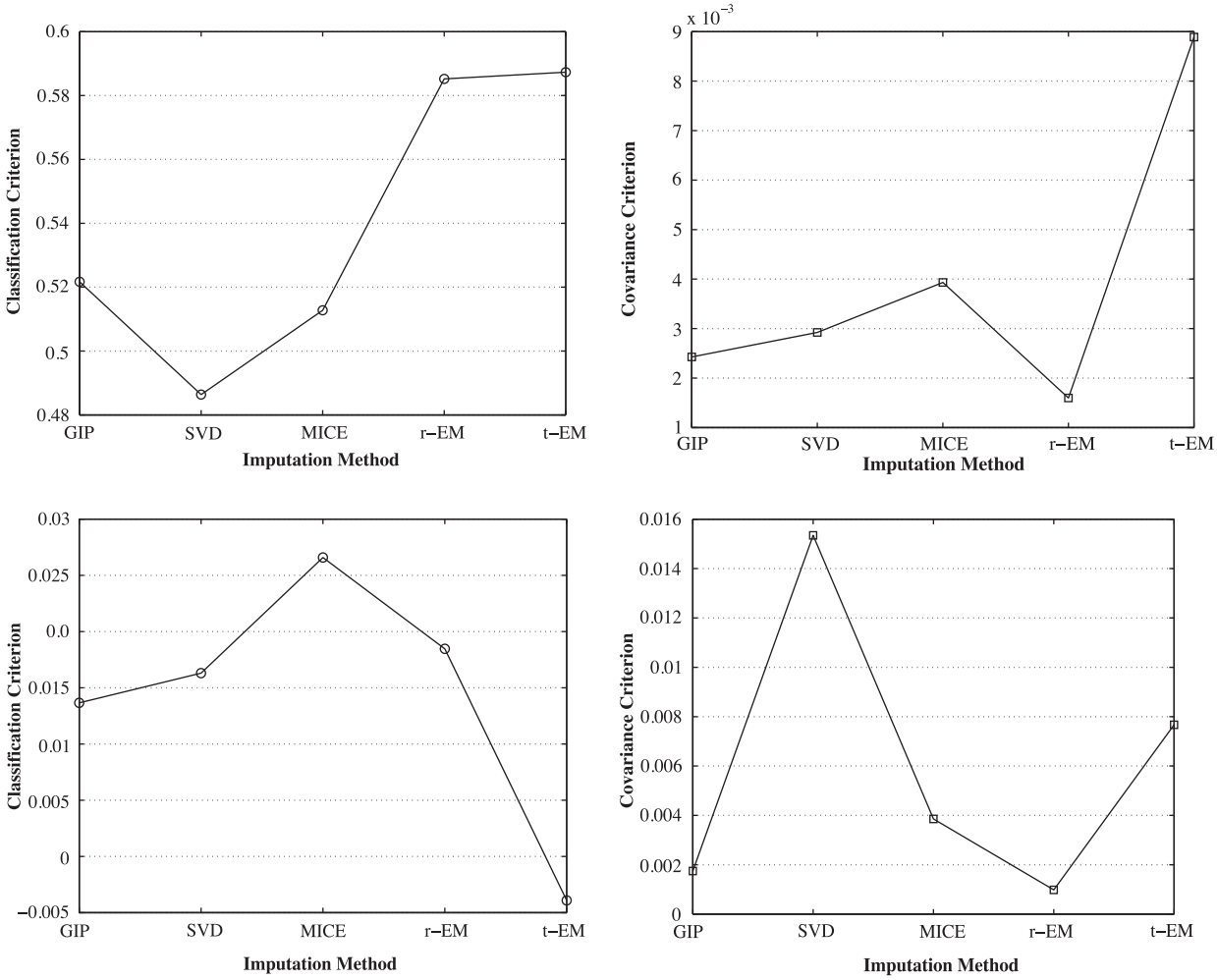

Fig. 3. Comparison of different multivariate imputation methods via two criteria. a) classification criterion (ψ_{err}) for **M1**; b) covariance criterion (ψ_{cov}) for **M1**; c) classification criterion (ψ_{err})for **M2**; d) covariance criterion (ψ_{cov}) for **M2**.

indicated in Fig. 2. In addition, because the ratio of diagonal elements of A_k from **X** is closer to 1 than the ratios from **G1** and **G2**, the dashed ellipse that represents the covariance matrix of **X** in Fig. 2 is more circular than the ellipses that represent the covariance matrixes for **G1** and **G2**. The two variables in the simulated dataset as well as in each class are moderately correlated, as seen in Table 1.

The covariance structures of the imputed datasets were compared with the ones of the complete dataset. Two typical imputation methods are selected for each dataset. These imputation methods are chosen because all of them show appreciable distinction between covariance criterion metrics and classification criterion metrics (cf. Fig. 3). In addition, other imputation methods for **M1** set and **M2** set reveal results similar to those for the ones chosen. For **M1**, MICE and r-EM are chosen. For convenience, **R11** refers to the complete dataset after imputation by MICE on the **M1** set and **R12** refers to the complete dataset after imputation by r-EM on the M1 set. For the **M2** set, SVD and r-EM imputation are chosen and the complete datasets represented as **R21** and **R22** for convenience. To investigate the covariance structure of the imputed datasets, covariance matrices of **R11**, **R12**, **R21** and **R22** are also decomposed based on Eq. (15). Results are shown in Table 2. The decomposition results show that the covariance matrices of all the complete, imputed datasets have the same eigenvectors $\mathbf{D}_k = \begin{pmatrix} 0.7010 & -0.7070 \\ 0.7071 & 0.7071 \end{pmatrix}$, which indicates that the orientation of the covariance does not change with imputation. Comparing λ_k after imputation in Table 2 with its value from the complete dataset shown in Table 1, the eigenvalue λ_k of **R22** has the minimum change from its original value, that from **R12** has a small

Table 1
Covariance structure of the simulated dataset.*

	$\mathbf{\Sigma}_k$	D_k	λ_k	A_k
X	[1.0000 0.3405 0.3405 1.0000]	[0.7071 −0.7071 0.7071 0.7071]	0.9402	[1.4258 0 0 0.7014]
G1	[1.0000 0.4840 0.4840 1.0000]	[0.7071 −0.7071 0.7071 0.7071]	0.8750	[1.6960 0 0 0.5896]
G2	[1.0000 0.4840 0.4840 1.0000]	[0.7071 −0.7071 0.7071 0.7071]	0.8750	[1.6960 0 0 0.5896]

* Parameters $\mathbf{\Sigma}_k$ D_k λ_k and A_k are defined in Eq. (15).

Table 2
Volume (λ_k) of covariance matrices for the imputed simulated dataset **X** and its class matrices **G1** and **G2**.

	M1 imputed by MICE (**R11***)	**M1** imputed by r-EM (**R12***)	**M2** imputed by SVD (**R21***)	**M2** imputed by r-EM (**R22***)
X	0.966	0.935	0.978	0.943
G1	0.914	0.872	0.937	0.922
G2	0.966	0.903	0.956	0.897

* The letters **R11**, **R12**, **R21**, and **R22** correspond to the datasets shown in Fig. 4a, b, c and d, respectively.

change, that from **R11** has a relatively large change and the eigenvalue λ_k from **R21** has the maximum change from the original value of λ_k: $R22 < R12 < R11 < R21$. This trend is the same as with the covariance change criterion shown in Fig. 3b and d. However, the covariance volume, which λ_k represents, does not correlate well with classification performance in this dataset. As shown in Fig. 3 and Table 2, all covariance criterion results, as well as the λ_k values, are of the same order of magnitude, no matter whether the missing elements occur in the first variable or in the second. However, it is obvious that the classification errors of the imputed datasets do not change much from those obtained on the complete dataset **X** only when the missing elements are located in the second variable. If missing elements appear at the first variable in this simulated dataset, the change in classification error, ψ_{err}, is about 100 times larger than the change in the classification error obtained for the imputed datasets where missing elements occur in the second variable. Even though the performance of the imputed datasets from the **M1** set and those from the **M2** set is quite different, r-EM imputation shows good control on the volume of the whole data covariance matrix as well as that of the class covariance matrix, no matter where missing elements appear, as seen by comparing λ_k in Tables 1 and 2. This result suggests that covariance structures affected by imputation do not depend on which variables have missing elements, but are more related to the specific imputation method used on the simulated dataset. This finding is different from what was seen from the classification criterion.

The changes seen in the classification criterion can be explained by visualizing the QDA classifier and by calculating Wilks' lambda as mentioned in Section 3.3. Following the scheme presented in Fig. 2, the covariance structure as well as the discriminant boundaries of **R11**, **R12**, **R21** and **R22** is shown in Fig. 4. The main difference between those four sub-figures is the discriminant boundary. This boundary does not change its shape obviously in Fig. 4a and b as compared with that shown in Fig. 2, but in Fig. 4c and d, the discriminant boundary becomes curved. Note that the directions of the boundaries are the same. Considering the classification errors obtained from these four datasets in Fig. 4, it is clear that changes in class boundary that occur in Fig. 4c and d improve classification performance, but do not prove that the changes in class boundary result from imputation-related changes in the covariance matrix. The covariance matrices deviate more after imputation for **R11** and **R12** than the covariance matrix for **R22**, but the discriminant boundaries for **R11** and **R12** (Fig. 4a and b) are closer to lines than those for **R21** and **R22**, which indicates that the changes in each group covariance matrix of **R11** and **R12** are smaller than those of **R21** and **R22**. Improvement in classification performance, which will lead to lower values of the classification criterion, shows a connection with the position of the missing elements.

The inconsistency of the two criteria can be explained in part by considering results from the Wilks' lambda variable selection method. Applying the method mentioned in Section 3.3 to this simulated dataset, the new order of variables $X_{new} = [x_1, x_2]$ is the same as the

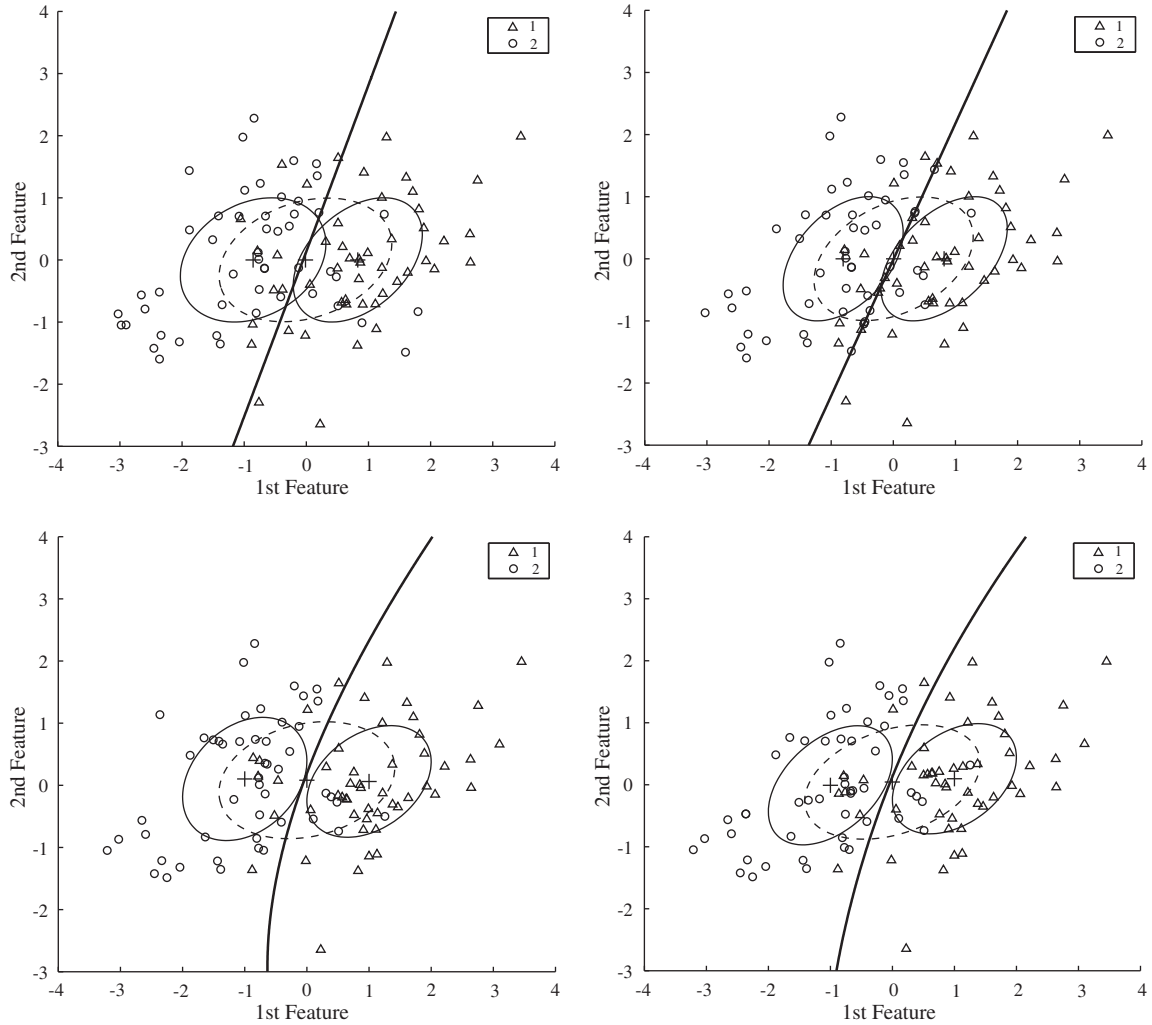

Fig. 4. Covariance structure and discriminant boundary of data after imputation. a) **M1** imputed by MICE (**R11**); b) **M1** imputed by r-EM (**R12**); c) **M2** imputed by SVD (R21); d) **M2** imputed by r-EM (**R22**). The meanings of all symbols are as given in Fig. 2.

Table 3
Imputation performance* on two incomplete versions of the Iris dataset.

	Covariance criterion					Classification criterion				
	GIP	SVD	MICE	r-EM	t-EM	GIP	SVD	MICE	r-EM	t-EM
3rd variable	254.43	60.43	520.36	4.68	3.95	0.51	0.49	1.12	0.46	0.46
1st variable	22.62	565.68	280.97	2.83	10.86	0.24	0.27	0.35	0.13	0.25

* "1st variable" refers to the dataset with all missing elements in the first variable. "3rd variable" refers to the dataset with all missing elements in the third variable.

original order, which indicates that the first variable has the most discriminant information. The Wilks' lambda for x_1 alone is 0.4949 and for the set $[x_1, x_2]$ the Wilks' lambda is 0.4287. This result indicates that both variables contribute to the classification, consistent with the correlation matrix $\mathbf{\Sigma}_k$ of $\mathbf{X}$ shown in Table 1. Despite the significance of the first two variables to the classifier, any changes to the first variable by imputation have more impact on classification performance than an imputation-based change to the second variable because the first variable has the larger discriminative power. That finding is consistent with the observation in Fig. 3 that **R11** and **R12** show higher classification errors compared with **R21** and **R22**.

Based on results from imputation of the simulated dataset, it is clear that changes in classification error depend strongly on the location of the missing elements. Not surprisingly, the classification error of the imputed dataset is higher when missing elements occur in the variable having more discriminative power, the first variable in the simulated dataset for instance, regardless of the choice of imputation method. On the other hand, changes in covariance structure are less affected by the position of missing elements. As a consequence, the two criteria are not consistent in identifying the best imputation method.

4.2. Iris dataset

The classification performance of the imputed datasets is shown above to depend on which variables are missing in the simulated dataset **X**. Even though the covariance matrix for the imputed dataset may be close to the covariance matrix for the original, complete dataset when the missing mechanism is MCAR, the classification performance based on the imputed dataset cannot be predicted from the estimated covariance parameters only. In this section, the well-known Iris dataset is used to demonstrate this point as an example with more than two variables.

The Iris dataset contains three classes, with each class having 50 samples. There are four variables in this dataset, sepal length, sepal width, petal length and petal width, respectively. These variables are not equally significant to separate the three classes. Based on Wilks' lambda variable selection, the new order of variables in the Iris dataset is $X_{new} = [x_3, x_2, x_4, x_1]$, which indicates that x_3 best discriminates the three classes. By combining other three variables sequentially with x_3, the Wilks' lambda values are 0.0586, 0.0369, 0.0250 and 0.0234. Because its value significantly decreases from 0.0369 to 0.0250 via combining x_4 to $[x_3, x_2]$, x_4 also shows an ability to provide better discrimination, but x_1 is not as significant as others because the Wilks' lambda value is not changed much by adding this variable.

As with the simulated dataset, missing elements in the Iris dataset are created randomly in variables with the most and least discriminant power, respectively, in order to examine the difference between the two criteria. 10% missing elements are created randomly in the 3rd variable to generate one incomplete dataset and then in the 1st variable to generate another dataset. The missing data structure of both datasets is MCAR. The five iterative imputation methods discussed above are then used to fill in missing values. Results from the two criteria are shown in Table 3. When missing elements occur only in the 3rd variable, the two criteria are consistent with each other; imputations that show higher ψ_{cov} also have higher ψ_{err}. However, that consistency is not the case when the missing elements occur only in the 1st variable. The ψ_{cov} value obtained after SVD imputation is higher than that for MICE imputation, but the value of ψ_{err} is lower. Furthermore, similar to what is seen in the simulated dataset, the change in classification error is smaller when the missing elements appear in less significant variables, in this case, in the 1st variable of the Iris dataset. For instance, the ψ_{err} metric values from GIP, SVD and t-EM imputation methods performed on the dataset in which only the 3rd variable has missing elements are almost two times higher than the ψ_{err} metric from the dataset in which only the 1st variable has missing elements. For the rest of the imputation methods, the difference between the two ψ_{err} metrics is even larger.

To expand this point to varied amounts of missing data in the Iris dataset, different proportions (5%, 10%, 15%, 20%, 25%, 30%, 35%, and 40%) of the data elements are deleted from the Iris dataset randomly to simulate sets with varying amounts of missing data. Unlike the

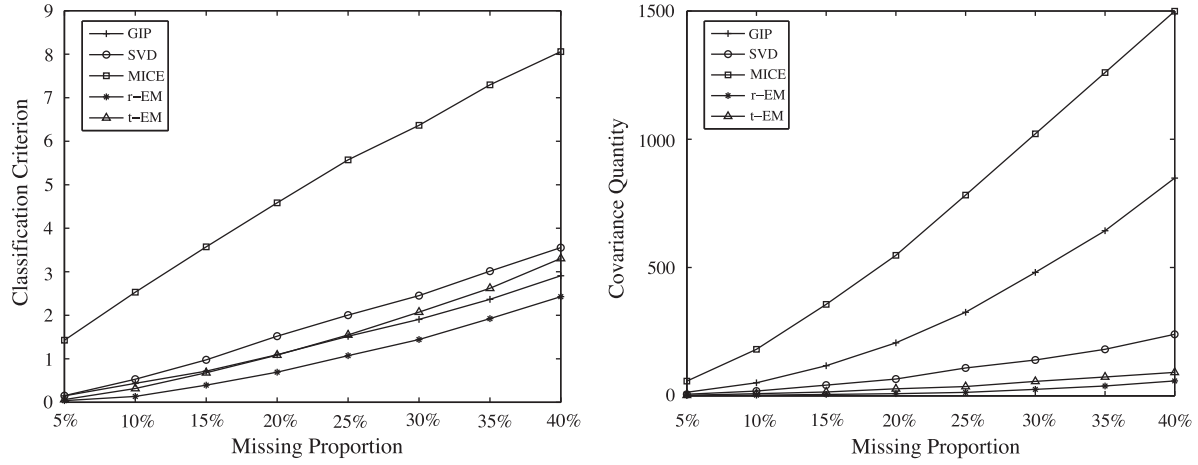

Fig. 5. Comparison of five imputation algorithms on the Iris dataset. (a) Classification criterion ψ_{err}; (b) covariance criterion ψ_{cov}.

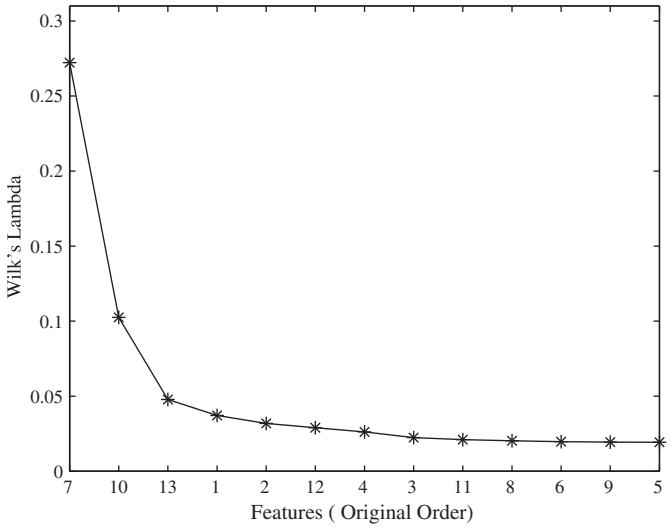

Fig. 6. Wilks' lambda analysis for the Wine dataset.

approach taken above, though, the missing elements are not restricted to specific variables, but are spread randomly in all variables. The missing mechanism is still MCAR because the distribution of missing values does not depend on the remaining data or the missing data when data elements are deleted randomly. The five imputation methods described above are then used on these incomplete datasets. After the complete datasets are obtained by means of the multivariate imputation, both imputation criteria described in Section 3 are calculated for all of the imputed data sets. The results are shown in Fig. 5. Comparing Fig. 5a and b, it is clear that both criteria increase as the proportion of missing data increases for any given imputation algorithm. In addition, for the five imputation methods examined here, the increase in the error of classification of an imputed data matrix is not always positively correlated with its change in covariance. For example, the GIP imputation algorithm shows smaller ψ_{err} than those associated with SVD in Fig. 5a, but the GIP algorithm has higher ψ_{cov} values than SVD in Fig. 5b. No direct relation between the two criteria can be found, like the situation when missing elements occur in only the 1st variable of the Iris dataset.

Even though the two criteria do not agree, it is not difficult to determine the best imputation method in this dataset because r-EM imputation shows the best performance (lowest values of ψ_{err} and ψ_{cov}) in both criteria.

In general, classification errors affected by imputation do not always relate with changes in covariance matrices. These errors are more dependent on which variable has missing elements. If missing elements appear in a variable having more discriminant power, the classification error and ψ_{err} are higher than those for missing values in a variable having less discriminant power. This conclusion is consistent with that from the simulated dataset.

4.3. Wine dataset

The effects of missing data on the Wine dataset were also examined. The total number of samples in this dataset is 178, with three classes containing 59, 71 and 48 samples, respectively. Similar to the Iris dataset, variables in the Wine dataset are not equally significant to discriminate classes. By using Wilks' lambda variable selection method as above, variables were rearranged according to their discriminant power. The Wilks' lambda result shown in Fig. 6 indicates that the 7th, 10th and 13th variables best distinguish the three classes, and because the value of Wilks' lambda does not change after the first three variables, adding more variables does not significantly change the ratio of the variances between classes and within classes. Based on the results from the simulated dataset and Iris dataset reported above, it can be inferred that ψ_{err} is higher when missing elements occur in features 7, 10 and 13 (the first three variables) than when missing data occur in other variables.

Varied amounts of missing data are again tested here. As with the Iris dataset, different proportions (5%, 10%, 15%, 20%, 25%, 30%, 35%, and 40%) of the data elements are deleted randomly to simulate sets with missing data. The missing mechanism is MCAR for reasons discussed above. Five imputation methods are then used to impute those missing values. Their performance is evaluated by calculation of the covariance and classification criteria and results are shown in Fig. 7. This figure shows some similarity with Fig. 5. The results are similar to those found with the Iris dataset, especially in the trends along the missing proportions: imputation performance deteriorates as the proportion of missing data increases. However, the consistency between the two criteria in the Wine dataset is different than that seen in the Iris dataset. The two criteria for the Iris dataset in Fig. 5 are consistent with each other in all imputation methods except for GIP imputation, but these criteria do not agree with each other for all of the imputation methods used on the Wine dataset. Based on the classification criterion alone, r-EM, t-EM, GIP, SVD and MICE range from the best imputation method to the worst one for the Wine dataset. However, based on only the covariance criterion, the order of

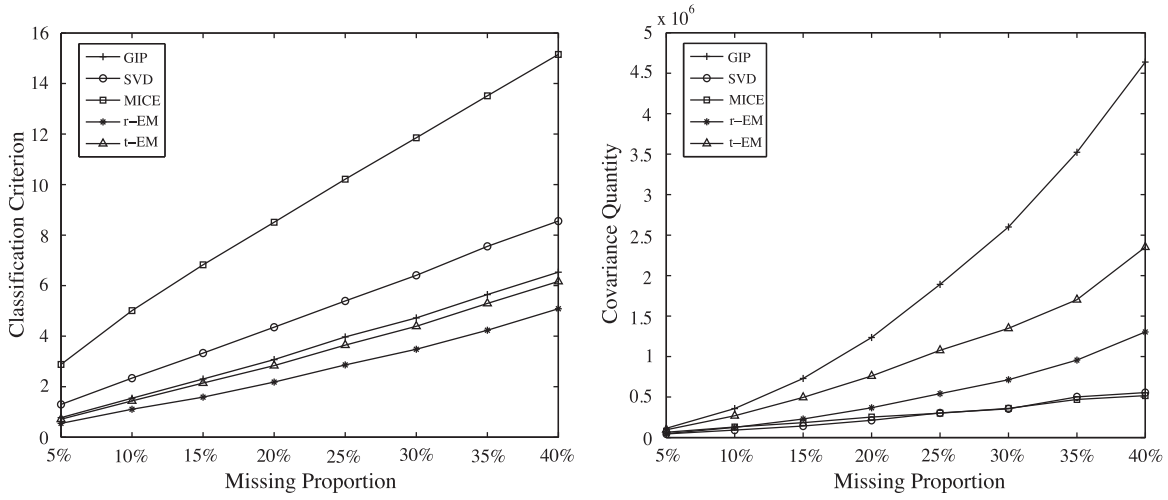

Fig. 7. Comparison of five imputation algorithms on the Iris dataset. (A) Classification criterion ψ_{err}; (B) covariance criterion ψ_{cov}.

imputation performance, from best to worst, is SVD, MICE, r-EM, t-EM and GIP, a sequence that is completely different. The inconsistency can be explained by difference in the ratio of variables that can discriminate classes. In the Iris dataset, three out of four variables can be determined as significant variables for the QDA classifier. When missing elements appear randomly in all variables, the two criteria tend to follow the relation seen in Table 3 when missing elements occur only in the most discriminating variable (3rd variable): classification performance is positively correlated with changes in the covariance matrix. On the contrary, in the Wine dataset, only three (7th, 10th, and 13th variables) out of thirteen variables are significant to the QDA classifier based on Wilks' lambda results. When missing elements appear randomly in all thirteen variables, the proportion of missing elements in those three significant variables is small. Results from the two criteria shown in Fig. 7 tend to follow the relation seen in Table 3, where missing elements occur only in the least significant variable (1st variable): classification performance is not correlated with changes in the covariance matrix.

For the Wine dataset, the best imputation method cannot be determined directly, as in the Iris dataset. SVD imputation provides the lowest ψ_{cov} but it has relatively high ψ_{err}. Imputation using the r-EM method has the lowest ψ_{err} but relatively high ψ_{cov}. No imputation method dominates in both criteria. Imputation with r-EM is a good choice when the missing proportion is lower than 15% because ψ_{cov} for r-EM imputation is close to the lowest value in this situation. At higher missing proportions, choosing the best imputation method depends on the requirements imposed by any further analysis after obtaining the complete dataset. If classification is the goal, the r-EM imputation method is suggested; for other uses of the data, SVD imputation may be a better choice.

5. Conclusions and suggestions

Based on the results seen from several data sets and five algorithms for iterative imputation, the covariance criterion ψ_{cov} does not always correlate with the classification criterion ψ_{err} after the imputation of missing data. When missing elements occur in variables that explain significant amounts of variance occurring between and within classes, the values of ψ_{cov} and ψ_{err} metrics appear to be positively correlated most of the time. On the contrary, the values of ψ_{cov} and ψ_{err} metrics appear to disagree with each other when missing elements occur in variables that explain less of the variance.

No single imputation method emerged as the overall best in all cases examined. The r-EM imputation method is competitive when the missing proportion is under 20%, judging from results obtained from both real datasets. If the percentage of the missing data is above 20%, the final purpose of analyzing practical datasets should be considered carefully before choosing an imputation method.

Of course, population covariance and class assignments are usually unknown for practical, incomplete data. In this situation, neither criterion discussed above can be used to compare the performance of the imputation. Besides, Wilks' lambda may provide incorrect variable selection based on the incomplete data. From the studies above, the r-EM imputation algorithm is a safe choice given its performance and its relative stability on all data sets examined. Nevertheless, imputation of a dataset containing a high proportion of missing data can lead to irreversible damage to the statistical properties of the data, which will make reliable imputation impossible.

Acknowledgements

Funding was provided by NGA-NURI though a research grant.

Datasets are obtained from Frank, A. & Asuncion, A. (2010, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.)

References

[1] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, second ed. Wiley-Interscience, New Jersey, 2002.
[2] J.L. Schafer, Analysis of Incomplete Multivariate Data, first ed. CRC Press, New York, 1997.
[3] T. Schneider, Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values, Journal of Climate 14 (2001) 853–871.
[4] B. Walczak, D. Massart, Dealing with missing data, part I, Chemometrics and Intelligent Laboratory Systems 58 (2001) 15–27.
[5] P.R.C. Nelson, P.A. Taylor, J.F. MacGregor, Missing data methods in PCA and PLS: score calculations with incomplete observations, Chemometrics and Intelligent Laboratory Systems 35 (1996) 45–65.
[6] R. Lopez-Negrete, S. García-Muñoz, L.T. Biegler, An efficient nonlinear programming strategy for PCA models with incomplete data sets, Journal of Chemometrics 24 (2010) 301–311.
[7] B. Grung, R. Manne, Missing values in principal component analysis, Chemometrics and Intelligent Laboratory Systems 42 (1998) 125–139.
[8] J. Camacho, A. Ferrer, Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: theoretical aspects, Journal of Chemometrics 26 (2012) 361–373.
[9] F. Arteaga, A. Ferrer, Dealing with missing data in MSPC: several methods, different interpretations, some examples, Journal of Chemometrics 16 (2002) 408–418.
[10] A. Smolinski, B. Walczak, J.W. Einax, Exploratory analysis of data sets with missing elements and outliers, Chemosphere 49 (2002) 233–245.
[11] F. Arteaga, A. Ferrer, Framework for regression-based missing data imputation methods in on-line MSPC, Journal of Chemometrics 19 (2005) 439–447.
[12] I. Stanimirova, S. Serneels, P.J. Van Espen, B. Walczak, Dealing with missing values and outliers in principal component analysis, Analytica Chimica Acta 581 (2007) 324–332.
[13] A.L. Bello, Imputation techniques in regression analysis: looking closely at their implementation, Computational Statistics & Data Analysis 20 (1995) 45–57.
[14] A.L. Bello, Choosing among imputation techniques for incomplete multivariate data — a simulation study, Communications in Statistics A Theory 22 (1993) 853–877.
[15] K.V. Branden, S. Verboven, Robust data imputation, Computational Biology and Chemistry 33 (2009) 7–13.
[16] S. Oh, D.D. Kang, G.N. Brock, G.C. Tseng, Biological impact of missing-value imputation on downstream analyses of gene expression profiles, Bioinformatics 27 (2011) 78–86.
[17] D. Williams, X. Liao, Y. Xue, L. Carin, B. Krishnapuram, On classification with incomplete data, IEEE Transactions on Pattern Analysis 29 (2007) 427–436.
[18] A. Farhangfar, L. Kurgan, J. Dy, Impact of imputation of missing values on classification error for discrete data, Pattern Recognition 41 (2008) 3692–3705.
[19] J.G. Ibrahim, M.H. Chen, S.R. Lipsitz, A.H. Herring, Missing-data methods for generalized linear models, Journal of the American Statistical Association 100 (2005) 332–346.
[20] M.J. Azur, E.A. Stuart, C. Frangakis, P.J. Leaf, Multiple imputation by chained equations: what is it and how does it work? International Journal of Methods in Psychiatric Research 20 (2011) 40–49.
[21] J. Wang, Data Mining: Opportunities and Challenges, Idea Group Pub., PA, 2003.
[22] R.E. Dear, A principal-component missing-data method for multiple regression models, reprint, System Development Corp., Santa Monica, CA, 1959.
[23] L.S. Chan, O.J. Dunn, Treatment of missing values in discriminant analysis. 1. Sampling experiment, Journal of the American Statistical Association 67 (1972) 473–477.
[24] W.J. Krzanowski, Missing value imputation in multivariate data using the singular value decomposition of a matrix, Biometrical Letters 25 (1988) 31–39.
[25] W.J. Krzanowski, Cross-validation in principal component analysis, Biometrics 43 (1987) 575–584.
[26] I.R. White, P. Royston, A.M. Wood, Multiple imputation using chained equations: issues and guidance for practice, Statistics in Medicine 30 (2011) 377–399.
[27] K.J. Lee, J.B. Carlin, Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation, American Journal of Epidemiology 171 (2010) 624–632.
[28] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, Series B Methods 39 (1977) 1–38.
[29] R.D. Fierro, G.H. Golub, P.C. Hansen, D.P. O'Leary, Regularization by truncated total least squares, SIAM Journal on Scientific Computing 18 (1997) 1223–1241.
[30] K. Varmuza, P. Filzmoser, Introduction to Multivariate Statistical Analysis in Chemometrics, first ed. CRC Press, New York, 2009.
[31] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed. Wiley-Interscience, New Jersey, 2000.
[32] H. Bensmail, G. Celeux, Regularized Gaussian discriminant analysis through eigenvalue decomposition, Journal of the American Statistical Association 91 (1996) 1743–1748.
[33] J.D. Banfield, A.E. Raftery, Model-based Gaussian and non-Gaussian clustering, Biometrics 49 (1993) 803–821.
[34] B.W. Flury, M.J. Schmid, A. Narayanan, Error rates in quadratic discrimination with constraints on the covariance matrices, Journal of Classification 11 (1994) 101–120.
[35] P. Leray, P. Gallinari, Feature selection with neural networks, Behaviormetrika 26 (1998) 145–166.
[36] A. El Ouardighi, A. El Akadi, D. Aboutajdine, Feature selection on supervised classification using Wilks lambda statistic, in: International Symposium on Computational Intelligence and Intelligent Informatics (ISCIII'07), 2007, pp. 51–55.
[37] F. Arteaga, A. Ferrer, How to simulate normal data sets with the desired correlation structure, Chemometrics and Intelligent Laboratory Systems 101 (2010) 38–42.