International Journal of Innovative Computing, Information and Control
Volume **16**, Number **5**, October **2020**

ICIC International ©2020 ISSN 1349-4198
pp. **1681–1696**

IMPUTATION USING THE SINGULAR VALUE DECOMPOSITION: VARIANTS OF EXISTING METHODS, PROPOSED AND ASSESSED

Sergio Arciniegas-Alarcón[1], Marisol García-Peña[2,*] and Wojtek Janusz Krzanowski[3]

[1]Departamento de Matemáticas, Física y Estadística
Facultad de Ingeniería
Universidad de La Sabana
Campus Puente del Común, Km. 7 Autopista Norte, Chía, Colombia
sergio.arciniegas@unisabana.edu.co

[2]Departamento de Matemáticas
Facultad de Ciencias
Pontificia Universidad Javeriana
Carrera 7 40-62, Bogotá, Colombia
*Corresponding author: luzmara@gmail.com

[3]Department of Mathematics
College of Engineering Mathematics and Physical Sciences
University of Exeter
Harrison Building, North Park Road, Exeter, EX4 4QF, UK
W.J.Krzanowski@exeter.ac.uk

Received January 2020; revised May 2020

Abstract. *Complete data matrices are required for some statistical analysis techniques, making imputation of missing data necessary in certain circumstances. The Krzanowski imputation system is based on singular value decomposition of a matrix and has no distributional or structural assumptions, but the system needs an imputation refining process through an iterative scheme. Two such iterative schemes already exist: expectation-maximization, Bro et al. and parity check, Arciniegas-Alarcón et al. The aim of this study is to present new variants of the basic method and to determine which iterative scheme produces the higher quality imputations. For this a simulation study was performed, and from incomplete matrices the quality of the imputations was assessed by estimating their uncertainty and by other criteria such as variance, bias and mean square error when a parameter of interest is considered. The best results were found using iterations with parity check and eliminating the singular values of the imputation equation.*
Keywords: Missing values, Singular value decomposition, Uncertainty, Imputation, Iterative computational scheme

1. **Introduction.** Imputation is a technique that replaces by plausible values the missing elements of a matrix, and thereby enables the application of statistical analyses that require complete data matrices. Several distinct variations of imputation can be identified: simple imputation (SI), multiple imputation (MI), and imputation based on statistical models that depend on unobserved latent variables. Maximum likelihood estimation of parameters in such models is conducted using an iterative method known as the expectation-maximization (EM) algorithm [3].

Some of the classic references that formally present these methods are Dempster et al. [4], Seber [5, 6], Srivastava and Carter [7], Rubin [8], Little and Rubin [9], Srivastava

DOI: 10.24507/ijicic.16.05.1681

[10], Srivastava and Dolatabadi [11], and van Buuren [12]. More recent descriptions of these methodologies that describe new developments such as free software, three-way and three-mode multivariate analysis, and Bayesian statistics can be found in Tian et al. [13], Anindita et al. [14], Murray [15], Muharemi et al. [16], Matsuda and Komaki [17] and van Ginkel et al. [18].

The aforementioned methods depend heavily on structural and distributional assumptions because they may need, for instance, a normal multivariate distribution or their efficiency may also depend on the mechanism that is assumed to underlie the missing values in the data set under study: values may be missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) [9, 19, 20].

Because of this dependence, applied researchers may prefer general imputation schemes that are free from any distributional or structural restriction. One such nonparametric method is the Krzanowski imputation system based on the singular value decomposition (SVD) of a matrix [21]. Currently (May of 2020) in Google Scholar, this system numbers 47 citations and in Arciniegas-Alarcón et al. [2] can be found some references related to the historical evolution of the method from its first presentation. Specifically, in Arciniegas-Alarcón et al. [2] it was proposed to apply the Krzanowski system to real data from multi-environmental experiments, but in this work no conclusive results were obtained as to which iterative scheme (EM iterations or parity check iterations) will best avoid convergence issues [1, 22].

Consequently, nine simple variants of the imputation equation proposed by Krzanowski in 1988 will now be tested, five of them will use EM-type iterations and the remaining four will use parity check iterations. Of these nine variants, two will be considered as "gold standard" and the remaining seven are new variants, that constitute the paper's contribution to data analysis and that to our best knowledge have not yet been tested in the statistical literature concerned with the chosen imputation system.

To delimit the research and determine the most efficient iterative scheme, a simulation study was performed according to the methodology proposed by Heydarbeygie and Ahmadi [23]. Thus, incomplete matrices were generated under the MCAR mechanism and from them was estimated the uncertainty of the imputation of each of the proposed variants, testing their significance using a nonparametric method. Additional criteria suggested by these authors were also used.

The outline of this paper is as follows. Section 2 presents a review on the existing studies about data imputation. Section 3 presents two updated versions of the Krzanowski imputation method using iterations with parity check and EM, which will be considered as "gold standard" as they have already been presented in previous work. Section 4 presents seven new variants of the Krzanowski system through changes in the corresponding imputation equations. Section 5 presents a numerical assessment of the new proposals comparing them with the "gold standard" methods through a simulation study. In Sections 6 and 7 the results and discussion are presented and finally in Section 8 the conclusions of the work with possible open lines of research are described.

2. **Related Work.**

Strategies to circumvent the missing data problem have been described in the literature, so we first present a short review. In 1977, Dempster et al. [4] proposed a general method to iteratively calculate estimates by maximum likelihood where incomplete data exist, due, for example, to missing observations. Since each iteration of the algorithm consists of a step of expectation, followed by a step of maximization, the algorithm was called EM [5, 6].

Little and Rubin [9] summarized computational calculations as follows: "The EM algorithm formalizes a relatively old ad hoc idea for handling missing data: 1) replace missing

values by estimated values, 2) estimate parameters, 3) re-estimate the missing values assuming the new parameter estimates are correct, 4) re-estimate parameters, and so forth, iterating until convergence". An application of the EM algorithm is found in multivariate statistical analysis, because if it is assumed that sampling comes from an exponential family, the multivariate normal distribution is a member of it [6]. A presentation of the EM algorithm and Newton-Raphson alternatives, under the assumptions of normality and MAR missing data mechanism, to obtain maximum likelihood estimators of mean vectors and covariance matrices from incomplete multivariate data can be found in Srivastava and Carter [7] and Srivastava [10]. Bayesian developments have also been presented with the algorithm, see details in Matsuda and Komaki [17].

Another alternative for analyzing incomplete data is multiple imputation – MI [8]. This involves three distinct steps: (i) Imputation: The missing values are estimated M times, generating M completed data sets (observed + imputed); (ii) Analysis: The M completed data sets are analyzed using appropriate statistical procedures for the problem at hand; (iii) Combination: The M separate sets of results are combined into one single inference. A recent description of the technique can be found in Murray [15].

MI can be applied for univariate and/or multivariate imputation. For univariate imputation, linear models can be used [11] and for the multivariate case, Markov Chain Monte Carlo (MCMC) or Fully Conditional Specification (FCS) algorithms [12,14]. Although the EM and MI algorithms may be two of the best options that the modern theory of missing data can provide, on certain occasions these methods do not guarantee the same quality in the results in non-normal distributions and/or with MNAR missing data mechanisms [18].

For this reason, alternative methods have recently been used in [16], for example, schemes that use kNN (k nearest neighbors) algorithms or through imputations based in random forest (predictors that consist of a collection of randomized regression trees) which are highly versatile because they are non-parametric systems and can be applied to continuous, categorized or mixed data. In order to contribute to the growth of non-parametric imputation methods, our work proposes new alternatives based on the general scheme of Krzanowski [21] which is now presented.

3. **Krzanowski Imputation System with Parity Check or EM Iterations.**

Method 1 (M1). The first method consists of an updated version of the imputation system of Krzanowski (1988) with some minor changes that improve its performance [2]. Consider a matrix $\mathbf{Y}(n \times p)$ with elements y_{ij} $(i = 1, \ldots, n;\ j = 1, \ldots, p)$ and $p > n$ (if $p < n$ the matrix should be first transposed). First, suppose there is just one missing value y_{ij} in $\mathbf{Y}$. Then, the ith row from $\mathbf{Y}$ is deleted and the SVD for the $((n-1) \times p)$ resulting matrix $\mathbf{Y}^{(-i)}$ is calculated as $\mathbf{Y}^{(-i)} = \overline{\mathbf{U}}\overline{\mathbf{D}}\overline{\mathbf{V}}^T$, $\overline{\mathbf{U}} = (\overline{u}_{sh})$, $\overline{\mathbf{V}} = (\overline{v}_{sh})$, $\overline{\mathbf{D}} = \left(\overline{d}_1, \ldots, \overline{d}_p\right)$. The next step is to delete the jth column from $\mathbf{Y}$ and obtain the SVD for the $(n \times (p-1))$ matrix $\mathbf{Y}_{(-j)}$ as $\mathbf{Y}_{(-j)} = \widetilde{\mathbf{U}}\widetilde{\mathbf{D}}\widetilde{\mathbf{V}}^T$, $\widetilde{\mathbf{U}} = (\widetilde{u}_{sh})$, $\widetilde{\mathbf{V}} = (\widetilde{v}_{sh})$, $\widetilde{\mathbf{D}} = \left(\widetilde{d}_1, \ldots, \widetilde{d}_{p-1}\right)$. The matrices $\overline{\mathbf{U}}$, $\overline{\mathbf{V}}$, $\widetilde{\mathbf{U}}$ and $\widetilde{\mathbf{V}}$ are orthonormal, while $\widetilde{\mathbf{D}}$ and $\overline{\mathbf{D}}$ are diagonal. Now, combining the two SVDs, $\mathbf{Y}^{(-i)}$ and $\mathbf{Y}_{(-j)}$, the imputed value is given by

$$\widehat{y}_{ij} = \sum_{h=1}^{H} \widetilde{u}_{ih} \left(\widetilde{d}_h \sqrt{\frac{p}{p-1}}\right)^{\frac{1}{2}} \overline{v}_{jh} \left(\overline{d}_h \sqrt{\frac{n}{n-1}}\right)^{\frac{1}{2}} \tag{1}$$

where H is the optimal number of SVD components as found by cross-validation adapted for missing data matrices and available in the R statistical environment [24]. In this work we used the "*bcv*" package, which has implemented cross-validation for incomplete matrices using an EM algorithm [25, 26] (https://github.com/patperry/r-bcv).

When there is more than one missing value, an iterative scheme is required as follows. Initially all missing values are replaced by their respective column means, giving a completed matrix $\mathbf{Y}$ and then the columns are standardized by subtracting m_j and dividing the result by s_j (where m_j and s_j represent the mean and the standard deviation of the jth column calculated only from the observed values). Using the standardized matrix, the imputation for each missing value is recalculated using Equation (1). Finally, the matrix $\mathbf{Y}$ is returned to its original scale, $y_{ij} = m_j + s_j\widehat{y}_{ij}$. Then, the process is iterated until stability is achieved in the imputations. In order to avoid convergence problems, a parity check should be done in each iteration by matching the sign of $\left(\widetilde{u}_{ih}\left(\widetilde{d}_h\sqrt{\frac{p}{p-1}}\right)^{\frac{1}{2}}\right)\left(\overline{v}_{jh}\left(\overline{d}_h\sqrt{\frac{n}{n-1}}\right)^{\frac{1}{2}}\right)$ in (1) to the sign of $u_{ih}d_hv_{jh}$ obtained from the SVD of the $\mathbf{Y}$ matrix for each $h = 1, \ldots, H$ [2, 22].

Method 2 (M2). It is possible to avoid the parity check by using an alternative expression for (1) following the results of Bro et al. [1]. They suggest updating the missing y_{ij} by the corresponding element of the matrix

$$\mathbf{S} = \left(\widetilde{\mathbf{U}}\left(\widetilde{\mathbf{U}}\right)^{+}\right)\mathbf{Y}\left(\overline{\mathbf{V}}\left(\overline{\mathbf{V}}\right)^{+}\right)^{T} \tag{2}$$

where $(\bullet)^+$ represents the Moore-Penrose generalized inverse. Note that for each missing observation a different $\mathbf{S}$ matrix will be calculated; the inclusion in the algorithm of (2) makes the imputation basically an expectation maximization (EM) operation [1, 2].

4. **Proposed Variants.** Next, seven new extensions of the Krzanowski imputation system are presented. The extensions consist of proposing variants in the imputation equation but maintaining two of the main characteristics of the system originally chosen: these seven new extensions can be used in any data set or database that can be written in a matrix form without depending on some probability distribution or any missing data mechanism. These characteristics permit the extensions to be applied in various areas of knowledge, for example, agriculture, medicine, marketing, and engineering, contributing in the area of non-parametric statistical methods to incomplete data.

Method 3 (M3) and Method 4 (M4). Given that our only interest is the Krzanowski imputation system, method 1 and method 2 can be considered in this paper as "gold standard" because they have been previously presented and assessed in the literature, but Krzanowski suggested that d_h can also be estimated independently by $\widetilde{d}_h$ or $\overline{d}_h$. To our knowledge, this option has not yet been assessed in method-related references, which leads us to consider two simple variants of the imputation Equation (1) in the iterative parity check scheme:

$$\widehat{y}_{ij} = \sum_{h=1}^{H}\widetilde{u}_{ih}\widetilde{d}_h\sqrt{\frac{p}{p-1}}\overline{v}_{jh} \tag{3}$$

$$\widehat{y}_{ij} = \sum_{h=1}^{H}\widetilde{u}_{ih}\overline{v}_{jh}\overline{d}_h\sqrt{\frac{n}{n-1}} \tag{4}$$

The use of (3) gives method 3 and the use of (4) gives method 4.

Method 5 (M5). The fifth variant eliminates $\widetilde{d}_h\sqrt{\frac{p}{p-1}}$ and $\overline{d}_h\sqrt{\frac{n}{n-1}}$ by replacing Equation (1) of the iterative parity check scheme with

$$\widehat{y}_{ij} = \sum_{h=1}^{H} \widetilde{u}_{ih}\overline{v}_{jh} \tag{5}$$

This modification is based on the results of the research by Eshghi [27], who considered the effect of eigenvalues to be negligible when cross-validation is used for the choice of the optimal number of dimensions in the principal component analysis (PCA) from incomplete matrices. In the Eshghi study, a modified version of the Eastment and Krzanowski [22] cross-validation method was proposed, but in place of the SVD for the component model adjustment, the nonlinear iterative partial least squares (NIPALS) algorithm was used and had the disadvantage that the number of missing data could not exceed 20%. Eshghi's [27] justification for proposing this modification is based on the fact that the original method is computationally very intensive as the size of the matrix under analysis increases, because it requires the fit of two PCA models for each element of the matrix that is predicted. In addition, this approach is asymptotically inconsistent, which does not always lead to the selection of the model with the best predictive capacity (for details on real problems solved using predictive models see Gunawan [28]). The inconsistency can be corrected using a leave-group-out cross-validation approach, a fact that is used in the following proposed methods.

Method 6 (M6) and Method 7 (M7). In methods traditionally used for cross-validation of PCA, such as those proposed by Eastment and Krzanowski [22] and Gabriel [29], only one element of the matrix is left out and is later predicted. Owen and Perry [30] and Eshghi [27] have recently proposed generalizations of these methods, in which submatrices or groups of rows (columns) can be left out to make the corresponding prediction. For example, Owen and Perry [30] recommended leaving out submatrices of dimension (2×2) and (3×3).

Taking account of these results it is proposed to modify method 2 by eliminating two rows and two columns, i.e., replacing Equation (2) with

$$\mathbf{S}_{(-\mathbf{2})} = \left(\widetilde{\mathbf{U}}_{(-2)}\left(\widetilde{\mathbf{U}}_{(-2)}\right)^{+}\right)\mathbf{Y}\left(\overline{\mathbf{V}}_{(-2)}\left(\overline{\mathbf{V}}_{(-2)}\right)^{+}\right)^{T} \tag{6}$$

where $\widetilde{\mathbf{U}}_{(-2)}$ represents the SVD matrix of $\mathbf{Y}$ found after eliminating two columns, namely $\mathbf{Y}_{(-2\ columns)} = \widetilde{\mathbf{U}}_{(-2)}\widetilde{\mathbf{D}}_{(-2)}\widetilde{\mathbf{V}}^{T}_{(-2)}$ and $\overline{\mathbf{V}}_{(-2)}$ represents the SVD matrix of $\mathbf{Y}$ found after eliminating two rows, namely $\mathbf{Y}^{(-2\ rows)} = \overline{\mathbf{U}}_{(-2)}\overline{\mathbf{D}}_{(-2)}\overline{\mathbf{V}}^{T}_{(-2)}$.

Remember that in this EM algorithm, the missing y_{ij} is replaced by the corresponding value of $\mathbf{S}_{(-\mathbf{2})}$, so the first row eliminated from $\mathbf{Y}$ will be the ith and the second row will be randomly chosen from $(n-1)$ rows remaining. Similarly, the first column eliminated from $\mathbf{Y}$ will be the jth and the second column will be randomly chosen from $(p-1)$ columns remaining.

It is possible to consider eliminating three rows and three columns by replacing Equation (2) with

$$\mathbf{S}_{(-\mathbf{3})} = \left(\widetilde{\mathbf{U}}_{(-3)}\left(\widetilde{\mathbf{U}}_{(-3)}\right)^{+}\right)\mathbf{Y}\left(\overline{\mathbf{V}}_{(-3)}\left(\overline{\mathbf{V}}_{(-3)}\right)^{+}\right)^{T} \tag{7}$$

In this case, the first row eliminated from $\mathbf{Y}$ will be the ith and the other two rows will be chosen randomly from the remaining $(n-1)$ rows. Similarly, the first column eliminated from $\mathbf{Y}$ will be the jth and the other two will be randomly chosen from the remaining $(p-1)$ columns. EM algorithms using Equations (6) and (7) will be called M6 and M7 respectively.

Method 8 (M8) and Method 9 (M9). Two variants can be obtained from method 6 and method 7 described above. Both M6 and M7 eliminate the ith row and jth column, then randomly delete more rows (columns) depending on the size of the row group (columns) chosen. In this study we limited it to 2 and 3. However, given that the Krzanowski imputation method can be applied to any matrix, in a specifically multivariate data matrix [31] the rows and columns should be treated differently. Thus, if the rows represent independent individuals, random row deletion is reasonable after the obligatory deletion of the ith row, but in the columns there will generally be correlated variables. For this reason, we propose that in addition to the obligatory elimination of the jth column, the elimination of any remaining columns should be based on Spearman's correlation coefficient [32].

For example, if the goal is to impute the element y_{ij} and the chosen number of columns eliminated is 3, we suggest imputation in two phases. First apply the EM algorithm described in method 2 on the incomplete matrix $\mathbf{Y}$, and then on the completed matrix (observed + imputed) obtain the Spearman correlation matrix. The jth column and the two columns that have the least correlations with it are eliminated from $\mathbf{Y}$ to apply method 7. A similar procedure applies if only two columns are eliminated, but method 6 is applied. The application of method 6 and method 7 with the Spearman correlation criterion for elimination of columns is denoted by M8 and M9 respectively.

5. **Simulation Study.** To compare the variants of the Krzanowski imputation system, a simulation study was performed following the methodology proposed by Heydarbeygie and Ahmadi [23]. Thus, 1000 matrices $\mathbf{Z}$ of size (100×8) were generated with elements obtained from a uniform distribution $U(0, 1)$. For each matrix, 15%, 25% and 35% of the values were deleted randomly and treated as missing values, obtaining a total of 3000 incomplete matrices, denoted by $\mathbf{Z}_{miss}$. Subsequently, the nine variants of the Krzanowski system were applied to each incomplete matrix, obtaining completed matrices (observed + imputed) denoted by $\mathbf{Z}_{(completed)}$.

To compare the methods, the imputation uncertainty was estimated through a non-parametric approximation, obtaining a measure of its statistical significance. The procedure is described below. From the $\mathbf{Z}_{miss}$ matrix, 5% of the observed values of each column were eliminated [33], but initially recorded to compare them after imputation through the equation.

$$RD_{ij} = |z_{ij.trueval} - z_{ij.imputedval}|$$

where $z_{ij.trueval}$ is the true value that was eliminated and $z_{ij.imputedval}$ is the value produced by the imputation process. The RD_{ij} values are then ranked in ascending order and replaced by their ranks (i.e., the smallest becomes 1 and so on), and the Wilcoxon signed-rank test statistic is constructed:

$$T = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} RD_{ij}}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} RD_{ij}^2}}$$

where m and n are the number of rows and columns of $\mathbf{Z}_{miss}$ respectively. The significance measure of imputation uncertainty is approximately estimated by the p-value that is produced by the Wilcoxon test. The procedure must be repeated several times so that the variation of p-values can be considered. In this study, the procedure was repeated 10 times as in the studies by Solomon et al. [34].

Having obtained the p-values it is possible to establish if there is little or much uncertainty within the set of imputed values: at significance level α, p-values lower than α indicate high uncertainty in the imputations. In our study we used $\alpha = 5\%$ and from

the ten p-values obtained for each of the three thousand matrices, the distribution of the minimum p-value was studied. For additional details on this methodology, see [23].

One of the important steps of the previous procedure was repetition. Thus, for each incomplete $\mathbf{Z}_{miss}$ matrix, ten different $\mathbf{Z}_{(completed)}$ matrices were produced in each of which a parameter of interest was estimated and then the ten estimates were compared with the parameter estimate obtained from the corresponding original matrix $\mathbf{Z}$ without missing data. The comparison was made using mean bias and mean squared error (MSE). The ten estimates also made it possible to calculate variance as a criterion for selecting the most efficient imputation method. The parameter studied was the square root of the maximum eigenvalue of $\mathbf{Z}^T\mathbf{Z}$ (or equivalently, the maximum singular value obtained by the SVD of $\mathbf{Z}$), which in practice is used for biplot analysis or for analyses involving the additive main effects with the multiplicative interaction model AMMI [35, 36].

6. **Results and Analysis.** To establish whether any differences exist between our nine methods, and if any such differences do exist then to pinpoint which methods differ from the others, we consider the analysis of the data under four general headings: i) imputation

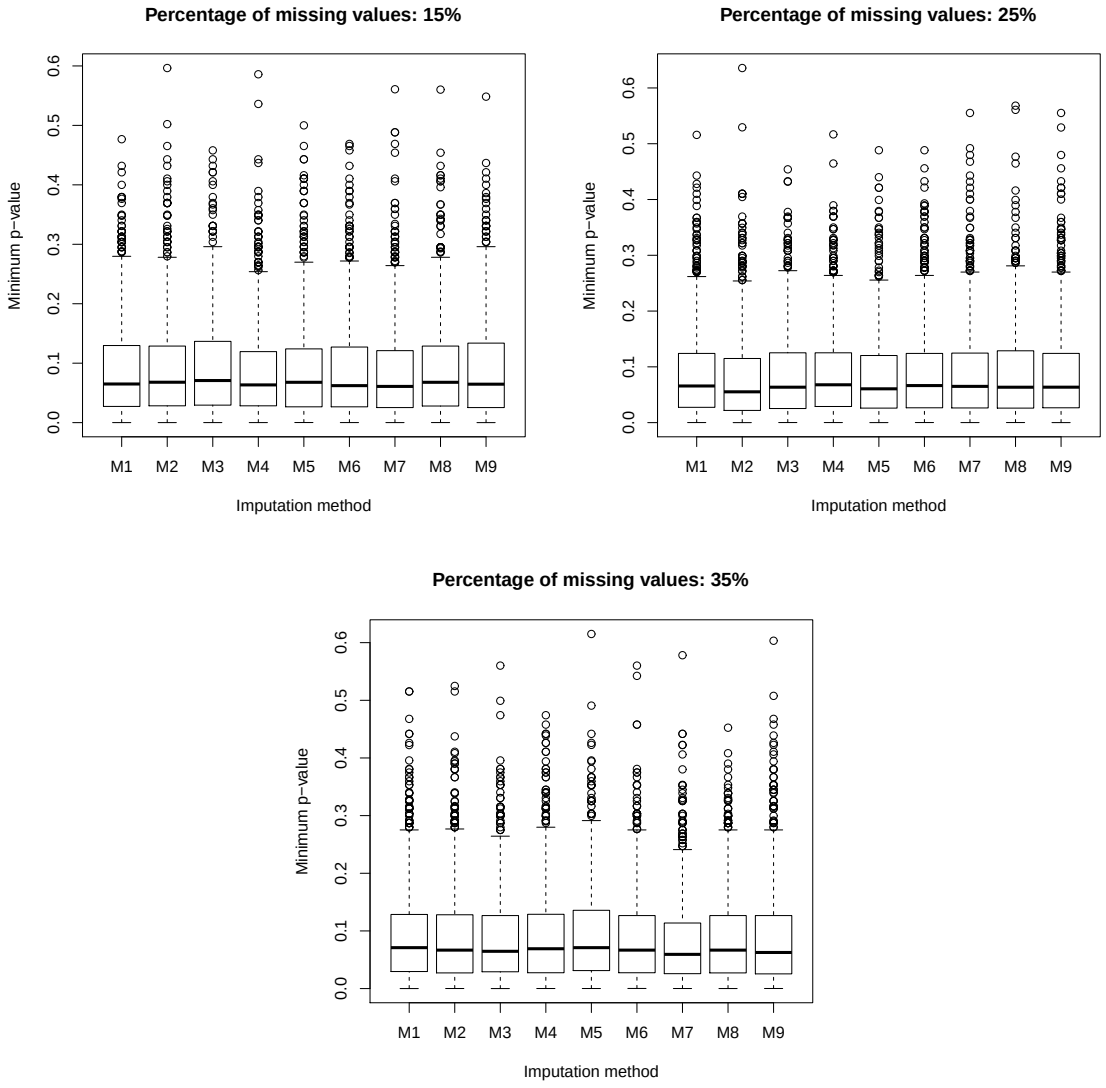

FIGURE 1. Distribution of minimum p-values in the different percentages of missing values

uncertainty of each method, ii) mean bias of each method, iii) mean squared error of each method, and iv) variance of the maximum singular value estimate. We now consider the detailed results for each of these headings in turn, and in each case we have a diagram and a table for each percentage of missing values in the data set.

6.1. **Imputation uncertainty.** The statistic used to test the significance of this uncertainty is the minimum p-value of each method for each simulation, and Figure 1 shows the distributions of these minimum p-values. An asymmetric distribution on the right is observed in the three percentages of missing values considered for all imputation systems. To compare the distributions, the nonparametric Friedman test [32] was performed, being significant only when 25% of the values were eliminated (p-value = 0.0324). This indicates that the uncertainty does not differ between the nine methods when considering the percentages of missing values of 15% (p-value = 0.1615) and 35% (p-value = 0.0625). After the only significant Friedman test, Wilcoxon-Nemenyi-MCDonald-Thompson's multiple

TABLE 1. Statistics of the minimum p-values

Percentage of missing values: 15%				
Method	Mean	Stdev	Median	IQR
M1	0.09010	0.08178	0.06498	0.10192
M2	0.09175	0.08453	0.06796	0.10032
M3	0.09370	0.08331	0.07080	0.10630
M4	0.08555	0.07870	0.06345	0.09100
M5	0.08891	0.08273	0.06785	0.09745
M6	0.08786	0.08155	0.06222	0.09974
M7	0.08539	0.08107	0.06091	0.09576
M8	0.08997	0.08246	0.06785	0.10081
M9	0.09013	0.08270	0.06466	0.10843
Percentage of missing values: 25%				
M1	0.08734	0.08106	0.06570	0.09661
M2	0.08099	0.08074	**0.05524**	0.09301
M3	0.08618	0.07828	0.06351	0.09990
M4	0.09078	0.08141	**0.06785**	0.09624
M5	0.08357	0.07660	0.06062	0.09431
M6	0.08839	0.08195	0.06642	0.09745
M7	0.08827	0.08301	0.06498	0.09794
M8	0.08761	0.08106	0.06347	0.10277
M9	0.08990	0.08618	0.06356	0.09747
Percentage of missing values: 35%				
M1	0.09330	0.08504	0.07098	0.09872
M2	0.08981	0.08344	0.06670	0.10083
M3	0.08840	0.08057	0.06465	0.09750
M4	0.09174	0.08533	0.06910	0.10125
M5	0.09556	0.08446	0.07098	0.10456
M6	0.08839	0.08107	0.06670	0.09917
M7	0.08311	0.07759	0.05935	0.08756
M8	0.08969	0.08050	0.06670	0.09945
M9	0.08930	0.08617	0.06263	0.10114

Methods that had a significant difference in nonparametric multiple comparisons are shown in bold.

comparisons were applied [32, 37] and it was found that only a significant difference existed between the M2 and M4. To observe the magnitude of the minimum p-values, the corresponding basic statistics are presented in Table 1, and it is observed that the mean and median of these p-values are greater than $\alpha = 5\%$, which means that the uncertainty was not significant in the imputed data for all methods in all percentages considered. With this criterion, it was found that the Krzanowski imputation system and its various variants produce very good quality imputations (non-significant uncertainty), both with the iterative scheme that includes parity check and iterations type EM.

6.2. **Mean bias.** This criterion is used to examine the accuracy of estimation of the maximum singular value of each simulated matrix, and Figure 2 shows the distributions of the mean bias for each imputation method considered. The scale of the box charts indicates that in the three percentages of missing data, the mean bias was negative, meaning that the maximum singular value of each original matrix simulated was overestimated when using the completed data (observed + imputed). This mean bias increased as the number of imputations increased, which is as expected. Taking this into account, the method

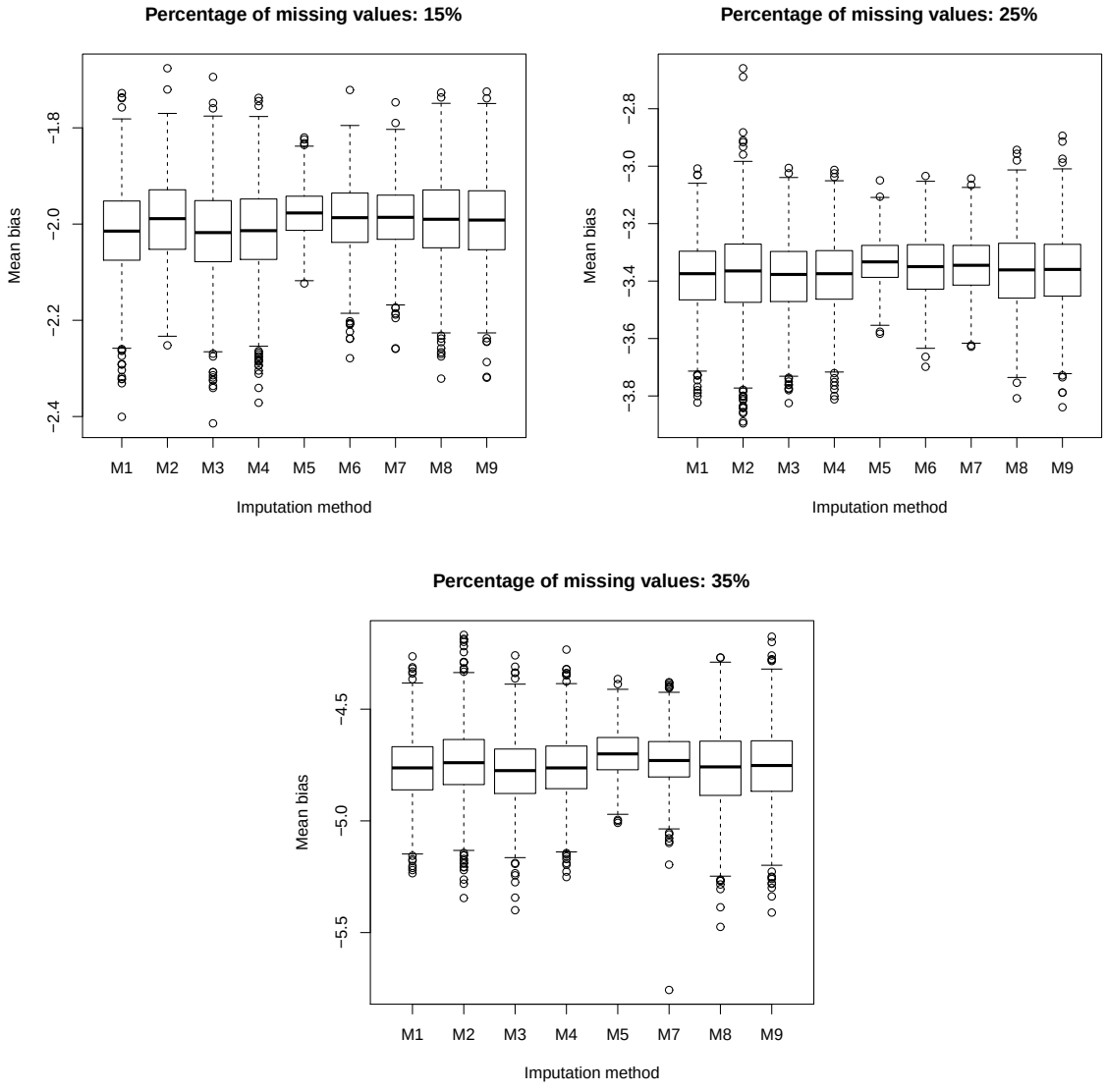

FIGURE 2. Distribution of mean bias in the different percentages of missing values

TABLE 2. Statistics of the mean bias

Method	Mean	Stdev	Median	IQR
Percentage of missing values: 15%				
M1	−2.01627	0.09735	−2.01466	0.12296
M2	−1.99057	0.08670	−1.98875	0.12362
M3	−2.01866	0.09926	−2.01774	0.12683
M4	−2.01562	0.09610	−2.01375	0.12550
M5	**−1.97786**	**0.05297**	**−1.97692**	**0.07089**
M6	−1.98799	0.07645	−1.98672	0.10240
M7	−1.98663	0.06958	−1.98604	0.09174
M8	−1.99158	0.09051	−1.99003	0.12020
M9	−1.99251	0.08864	−1.99169	0.12196
Percentage of missing values: 25%				
M1	−3.38173	0.13100	−3.37424	0.16938
M2	−3.37395	0.16470	−3.36470	0.20245
M3	−3.38703	0.13368	−3.37695	0.17433
M4	−3.38015	0.12939	−3.37427	0.16887
M5	**−3.33184**	**0.08069**	**−3.33318**	**0.11114**
M6	−3.35181	0.10770	−3.34968	0.15500
M7	−3.34580	0.09977	−3.34506	0.13818
M8	−3.36518	0.13412	−3.36107	0.19088
M9	−3.36285	0.13521	−3.35948	0.18025
Percentage of missing values: 35%				
M1	−4.76223	0.15083	−4.76284	0.19317
M2	−4.73387	0.16507	−4.73958	0.20146
M3	−4.77648	0.15494	−4.77499	0.19847
M4	−4.76057	0.14919	−4.76293	0.19081
M5	**−4.70124**	**0.10671**	**−4.70013**	**0.14417**
M7	−4.72430	0.12967	−4.72986	0.15833
M8	−4.76437	0.18517	−4.75829	0.24174
M9	−4.75280	0.17725	−4.75214	0.22599

Methods with the lowest statistics of the mean bias are shown in bold.

with the lowest mean bias can be considered the best and from a graphical point of view M5 stands out, the variant of the Krzanowski imputation system that uses parity check eliminating the singular values from the imputation equation. It can be seen graphically and in Table 2 that M5 in all percentages has the smallest variation (lowest standard deviations and lowest interquartile distances) and the mean and median closest to zero, i.e., with the smallest bias. Note that with 35% imputations M6 was eliminated from the box plot because M6 showed extremely large mean bias values not comparable with the remaining imputation systems.

The difference in mean bias between the methods was verified by the Friedman test and was significant in the three percentages considered, with p-values lower than 0.0001. Subsequently, the nonparametric multiple comparisons of Wilcoxon-Nemenyi-MCDonald-Thompson (WNMT) were performed. When 15% of the values were imputed, the centrality parameter of the M5 method differed from all remaining imputation systems except M6 (p-value = 0.0851) and M7 (p-value = 0.4574). When 25% of values were imputed, the only non-significant difference was found with the M7 method (p-value = 0.5488), and

when 35% of the values were imputed M5 had significant differences with all other methods (p-values < 0.0001). This analysis confirms that following the mean bias criterion, the recommended method is M5 (iterative parity check scheme) because it had significant differences with the other methods and when it had a centrality parameter equal to M6 and M7, it categorically outperformed them by its low dispersion (Table 2).

6.3. **Mean squared error.** This criterion is used to examine the overall spread of estimated parameter of each completed $\mathbf{Z}_{(completed)}$ matrix with respect to its corresponding original $\mathbf{Z}$ matrix, the parameter studied was the maximum singular value obtained by the SVD of $\mathbf{Z}$. Figure 3 and Table 3 display the values obtained for this criterion.

Looking at Figure 3 and Table 3, it can be established that in general, according to the MSE, again the M5 method is the most efficient with the lowest means, medians, standard deviations and interquartile distances. The Friedman test was performed to compare MSE distributions and was significant (p-values < 0.0001) in all the imputation percentages considered. Subsequently, with the multiple pair comparisons of WNMT, it was found

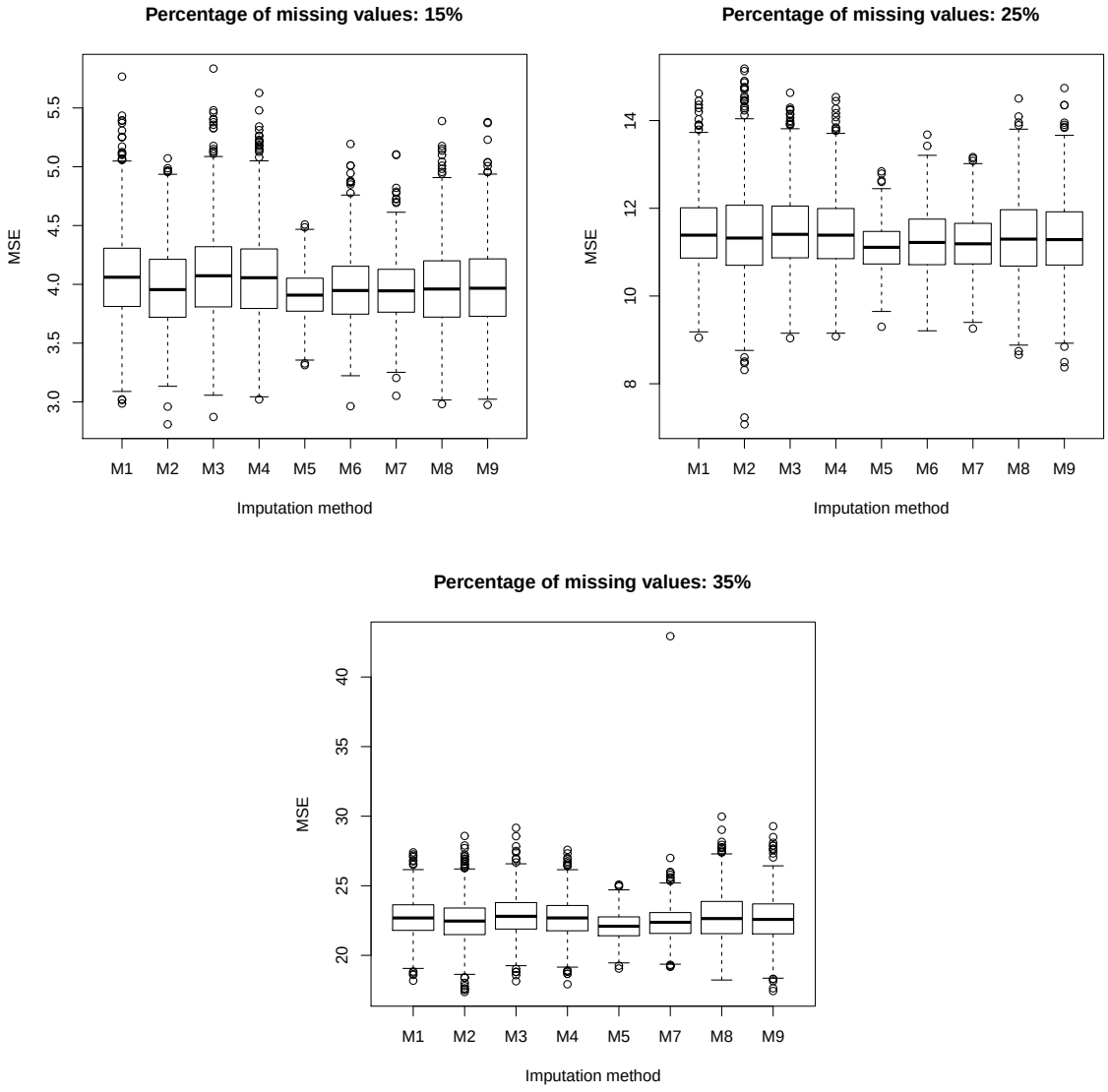

FIGURE 3. Distribution of mean squared error in the different percentages of missing values

TABLE 3. Statistics of the mean squared error – MSE

Method	Mean	Stdev	Median	IQR
Percentage of missing values: 15%				
M1	4.07583	0.39525	4.06040	0.49498
M2	3.97002	0.34589	3.95519	0.49211
M3	4.08588	0.40362	4.07299	0.51083
M4	4.07295	0.39001	4.05628	0.50515
M5	**3.91479**	**0.20957**	**3.90825**	**0.28036**
M6	3.95821	0.30508	3.94735	0.40700
M7	3.95181	0.27731	3.94474	0.36431
M8	3.97476	0.36201	3.96028	0.47820
M9	3.97813	0.35482	3.96704	0.48608
Percentage of missing values: 25%				
M1	11.45499	0.89019	11.38745	1.14537
M2	11.41152	1.11498	11.32184	1.36678
M3	11.49179	0.91012	11.40557	1.17900
M4	11.44388	0.87862	11.38744	1.14091
M5	**11.10771**	**0.53753**	**11.11011**	**0.74050**
M6	11.24688	0.72330	11.22092	1.03832
M7	11.20487	0.66874	11.19002	0.92435
M8	11.34300	0.90458	11.29713	1.28413
M9	11.32743	0.91077	11.28624	1.21270
Percentage of missing values: 35%				
M1	22.70414	1.43614	22.68605	1.83715
M2	22.43738	1.56298	22.46376	1.90922
M3	22.84197	1.48291	22.80492	1.89708
M4	22.68782	1.41944	22.68793	1.81622
M5	**22.11303**	**1.00313**	**22.09123**	**1.35487**
M7	22.34669	1.35228	22.37270	1.49564
M8	22.73659	1.76984	22.64380	2.30274
M9	22.62190	1.68654	22.58341	2.14959

The method with the lowest mean and median MSE is shown in bold.

that with the 15% imputation, the median MSE of the M5 method differs from other methods except M6 (p-value = 0.0814) and M7 (p-value = 0.4352). However, M5 remains preferable because it has the smallest variation (standard deviation and interquartile distance in Table 3). When considering the 25% imputation the only non-significant difference was detected with M7 (p-value = 0.5546), but looking at Table 3, again the MSE variation of M5 was the smallest (see variability measures). Finally, when 35% of the values were imputed, the centrality parameter of M5 turned out to be lower than the other methods (p-values < 0.0001). In summary, according to MSE the recommended method is the M5 method. Note that at 35% the M6 system did not present MSE results comparable with the other methods, so it was eliminated from the graph.

6.4. **Variance of the maximum singular value estimate.** This criterion is used to examine the spread of this estimate in each simulated matrix, and the M5 method again outperformed the remaining imputation systems. The efficiency of M5 is shown graphically in Figure 4 (the smallest variance in all imputation percentages). Nonparametric

analyses similar to those performed with the MSE and the mean bias were also performed with variance, but are not presented because the result was similar, indicating M5 as the best variant of Krzanowski's imputation system. Note that at 35% the M6 and M7 systems did not present MSE results comparable to the other methods; therefore, they were eliminated from Figure 4.

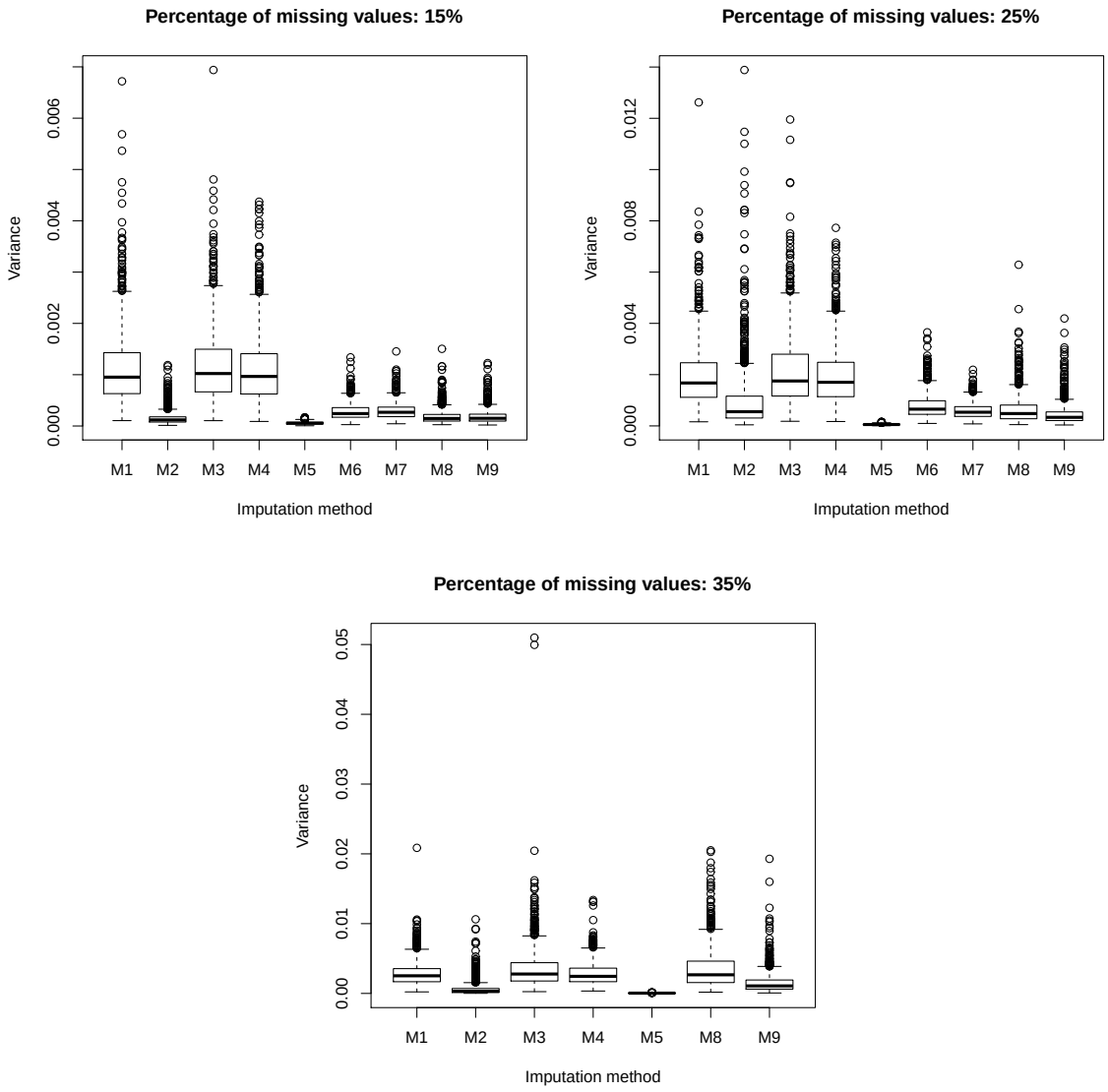

FIGURE 4. Distribution of variance of the maximum singular value in the different percentages of missing values

7. **Discussion.** The purpose of this research has been achieved. Nine variants of the Krzanowski imputation system [2, 21] were tested to determine which iterative scheme might be the most efficient. Variants based on iterations type EM (M6, M7, M8, and M9) depended on the M2 scheme by Bro et al. [1], but none of them outperformed the M5 variant that eliminated the singular values from the imputation equation and that works with iterations including parity check. In this study M1 (with parity check) and M2 (with EM iterations) were considered as the standard Krzanowski imputation systems, and these were largely outperformed by M5 using as criteria the mean bias, the MSE and the variance of the studied parameter.

Although the results are conclusive and relevant, they are not definitive because they leave open a line of research related to the imputation method by the singular value decomposition M5. Being a distribution-free method, it does not imply that it is robust to different probability distributions [38]; therefore, other simulations based on real data can be considered to test performance under different probability distributions other than the uniform one considered here [39, 40].

To delimit this research, only a missing completely at random data mechanism (MCAR) was considered to directly test the uncertainty of Krzanowski imputations with the nonparametric approach of Heydarbeygie and Ahmadi [23]. This methodology cannot be applied under other missing data mechanisms because its main element is the random elimination of elements, so it would be interesting to explore in future research nonparametric alternatives to estimate the uncertainty of the simple imputation of the M5 system under the MAR and MNAR mechanisms. While that research is conducted, when applied researchers are faced with incomplete data following these mechanisms, the performance of the M5 imputation method can be assessed with bias and/or MSE. Another way of estimating the uncertainty of imputations that does not depend on distributional or structural assumptions is by adapting the M5 variant to nonparametric multiple imputation following the schemes proposed by Bergamo et al. [41], Arciniegas-Alarcón et al. [2] and García-Peña et al. [42], this will undoubtedly need further research.

Finally, one aspect that may merit further research with the M5 imputation system is its performance when the matrices contain discrepant data, because SVD is a least squares technique that will necessarily be affected by such values. In this case, the first step is to detect them [43, 44] and later replace in the M5 the classic SVD with robust SVD [45], as proposed by Gabriel and Odoroff [46], Hawkins et al. [47], Liu et al. [48], Hernández-González and Galindo-Villardón [49], Jung [50], Zhang et al. [51] and Feng and He [52].

8. **Conclusions.** Of the seven new variants of the Krzanowski imputation system that were proposed in this study based on the decomposition by singular values, the best performance was obtained when the singular values of the imputation equation were eliminated within an iterative scheme using parity check. The proposal can be applied to any data set arranged in a matrix form and does not depend on distributional or structural assumptions, but it is assumed that the variables or columns of the incomplete matrix under study are correlated and without outliers. Future research may use the new proposal in studies of robustness to different probability distributions, robustness to different missing data mechanisms and in the assessment of resampling methods (bootstrap, jackknife or cross-validation) as tools to produce multiple imputation or to estimate uncertainty imputations in real data from any area of knowledge.

Acknowledgment. The authors of this paper acknowledge the High Performance Computing Center – ZINE of Pontificia Universidad Javeriana for assistance during the simulation study.

REFERENCES

[1] R. Bro, K. Kjeldahl, A. K. Smilde and H. A. L. Kiers, Cross-validation of component models: A critical look at current methods, *Anal. Bioanal. Chem.*, vol.390, pp.1241-1251, 2008.

[2] S. Arciniegas-Alarcón, M. García-Peña and W. J. Krzanowski, Missing value imputation in multi-environment trials: Reconsidering the Krzanowski method, *Crop Breed. Appl. Biot.*, vol.16, pp.77-85, 2016.

[3] W. Yan, Biplot analysis of incomplete two-way data, *Crop Sci.*, vol.53, pp.48-57, 2013.

[4] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. Royal Stat. Soc. Series B (Methodological)*, vol.39, no.1, pp.1-38, 1977.
[5] G. A. F. Seber, *Multivariate Observations*, John Wiley & Sons, New York, 1984.
[6] G. A. F. Seber, *Multivariate Observations*, John Wiley & Sons, New York, 2004.
[7] M. Srivastava and E. Carter, The maximum likelihood method for non-response in sample surveys, *Stat. Canada*, vol.12, pp.61-72, 1986.
[8] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York, 1987.
[9] R. Little and D. Rubin, *Statistical Analysis with Missing Data*, 2nd Edition, John Wiley & Sons, New York, 2002.
[10] M. S. Srivastava, *Methods of Multivariate Statistics*, John Wiley & Sons, New York, 2002.
[11] M. S. Srivastava and M. Dolatabadi, Multiple imputation and other resampling scheme for imputing missing observations, *J. Multivariate Anal.*, vol.100, pp.1919-1937, 2009.
[12] S. van Buuren, *Flexible Imputation of Missing Data*, CRC, Boca Raton, 2012.
[13] T. Tian, G. J. McLachlan, M. J. Dieters and K. E. Basford, Application of multiple imputation for missing values in three-way three-mode multi-environment trial data, *Plos One*, DOI: 10.1371/journal.pone.0144370, 2015.
[14] N. Anindita, H. A. Nugroho and T. B. Adji, A combination of multiple imputation and principal component analysis to handle missing value with arbitrary pattern, *The 7th International Annual Engineering Seminar (InAES)*, pp.1-5, 2017.
[15] J. S. Murray, Multiple imputation: A review of practical and theoretical findings, *Stat. Sci.*, vol.33, pp.142-159, 2018.
[16] F. Muharemi, D. Logofătu and F. Leon, *Review on General Techniques and Packages for Data Imputation in R on a Real-World Dataset*, Springer, 2018.
[17] T. Matsuda and F. Komaki, Empirical Bayes matrix completion, *Comput. Stat. Data Anal.*, vol.137, pp.195-210, 2019.
[18] J. R. van Ginkel, M. Linting, R. C. A. Rippe and A. van der Voort, Rebutting existing misconceptions about multiple imputation as a method for handling missing data, *J. Pers. Assess.*, pp.1-12, 2019.
[19] O. Harel and X. H. Zhou, Multiple imputation: Review of theory, implementation, and software, *Stat. Med.*, vol.26, pp.3057-3077, 2007.
[20] J. Paderewski and P. C. Rodrigues, The usefulness of EM-AMMI to study the influence of missing data pattern and application to Polish post-registration winter wheat data, *Aust. J. Crop Sci.*, vol.8, pp.640-645, 2014.
[21] W. J. Krzanowski, Missing value imputation in multivariate data using the singular value decomposition of a matrix, *Biomet. Lett.*, vol.25, pp.31-39, 1988.
[22] H. T. Eastment and W. J. Krzanowski, Cross-validatory choice of the number of components from a principal component analysis, *Technometrics*, vol.24, pp.73-77, 1982.
[23] A. Heydarbeygie and N. Ahmadi, Nonparametric methods for the estimation of imputation uncertainty, *J. Appl. Stat.*, vol.40, pp.693-698, 2013.
[24] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2019.
[25] P. O. Perry, *Cross-Validation for Unsupervised Learning*, Ph.D. Thesis, Stanford University, 2009.
[26] P. O. Perry, *bcv: Cross-Validation for the SVD (Bi-Cross-Validation)*, R package version 1.0.1., 2015.
[27] P. Eshghi, Dimensionality choice in principal components analysis via cross-validatory methods, *Chemom. Intell. Lab.*, vol.130, pp.6-13, 2014.
[28] F. E. Gunawan, Improving the reliability of F-statistic method by using linear support vector machine for structural health monitoring, *ICIC Express Letters*, vol.12, no.12, pp.1183-1193, 2018.
[29] K. R. Gabriel, The biplot – A tool for exploring multidimensional data, *J. Soc. Fr. Statistique*, vol.143, pp.5-55, 2002.
[30] A. Owen and P. Perry, Bi-cross-validation of the SVD and the nonnegative matrix factorization, *Ann. Appl. Stat.*, vol.3, pp.564-594, 2009.
[31] W. J. Krzanowski, *Principles of Multivariate Analysis: A User's Perspective*, Oxford University Press, Oxford, 2000.
[32] M. Hollander, D. A. Wolfe and E. Chicken, *Nonparametric Statistical Methods*, 3rd Edition, John Wiley & Sons, New York, 2014.
[33] N. Solomon, *A Stochastic Method for Estimating Imputation Accuracy*, Ph.D. Thesis, University of Sunderland, 2008.

[34] N. Solomon, G. Oatley and K. McGary, A dynamic method for the evaluation and comparison of imputation techniques, *Proc. of the World Congress on Engineering (WCE 2007)*, London, UK, 2007.
[35] H. G. Gauch, A simple protocol for AMMI analysis of yield trials, *Crop Sci.*, vol.53, pp.1860-1869, 2013.
[36] W. A. Malik and H. P. Piepho, Biplots: Do not stretch them!, *Crop Sci.*, vol.58, pp.1-9, 2018.
[37] D. G. Pereira, A. Afonso and F. M. Medeiros, Overview of Friedman's test and post-hoc analysis, *Communications in Statistics – Simulation and Computation*, vol.44, no.10, pp.2636-2653, 2015.
[38] A. L. Bello, Choosing among imputation techniques for incomplete multivariate data: A simulation study, *Commun. Stat.*, vol.22, pp.853-877, 1993.
[39] J. Forkman and H. P. Piepho, Robustness of the simple parametric bootstrap method for the additive main effects and multiplicative interaction (AMMI) model, *Biuletyn Oceny Odmian*, vol.34, pp.11-18, 2015.
[40] W. A. Malik, S. Hadasch, J. Forkman and H. P. Piepho, Non-parametric resampling methods for testing multiplicative terms in AMMI and GGE models for multi-environment trials, *Crop Sci.*, vol.58, pp.752-761, 2018.
[41] G. C. Bergamo, C. T. S. Dias and W. J. Krzanowski, Distribution-free multiple imputation in an interaction matrix through singular value decomposition, *Sci. Agr.*, vol.65, pp.422-427, 2008.
[42] M. García-Peña, S. Arciniegas-Alarcón, W. J. Krzanowski and D. Barbin, Multiple imputation procedures using the GabrielEigen algorithm, *Commun. Biomet. Crop Sci.*, vol.11, pp.149-163, 2016.
[43] F. Novika, Siswadi and T. Bakhtiar, The use of biplot analysis and Euclidean distance with Procrustes measure for outliers detection, *Int. J. Eng. Man. Res.*, vol.8, pp.194-200, 2018.
[44] V. Todorov, M. Templ and P. Filzmoser, Detection of multivariate outliers in business survey data with incomplete information, *Adv. Data Anal. Classif.*, vol.5, no.1, pp.37-56, 2011.
[45] P. C. Rodrigues, A. Monteiro and V. M. Lourenço, A robust AMMI model for the analysis of genotype-by-environment data, *Bioinformatics*, vol.32, pp.58-66, 2016.
[46] K. R. Gabriel and L. Odoroff, Resistant lower rank approximation of matrices, in *Data Analysis and Statistics III*, E. Diday et al. (eds.), Amsterdam, North-Holland, 1984.
[47] D. M. Hawkins, L. Liu and S. S. Young, *Robust Singular Value Decomposition*, National Institute of Statistical Sciences Technical Report 122, 2001.
[48] L. Liu, D. W. Hawkins, S. Ghosh and S. S. Young, Robust singular value decomposition analysis of microarray data, *Proc. of the National Academy of Sciences of the USA*, vol.100, 2003.
[49] S. Hernández-González and M. P. Galindo-Villardón, BIPROB: A method to obtain a robust biplot, *Rev. Inv. Ope.*, vol.27, pp.287-299, 2006.
[50] K. M. Jung, Robust singular value decomposition based on weighted least absolute deviation regression, *Commun. Kor. Stat. Soc.*, vol.17, pp.803-810, 2010.
[51] L. Zhang, H. Shen and J. Z. Huang, Robust regularized singular value decomposition with application to mortality data, *Ann. Appl. Stat.*, vol.7, no.3, pp.1540-1561, 2013.
[52] X. Feng and X. He, Robust low-rank data matrix approximations, *Sci. China Math.*, vol.2, pp.189-200, 2017.