Gazi University Journal of Science
GU J Sci
29(4): 799-809 (2016)

Imputation and Deletion Methods Under The Presence of Missing Values and Outliers: A Comparative Study

Onur TOKA[1, ♠], Meral ÇETİN[1]

[1]*Hacettepe University, Department of Statistics, Beytepe, 06800, Ankara, Turkey*

Received: 15/06/2016 Revised:16/08/2016 Accepted: 25/08/2016

ABSTRACT

Missing data and imputation methods are studied in many disciplines. However, the methods have some different properties and some constraints according to missingness mechanism. In this paper, we examine some deletion and imputation methods' behaviors under the presence of outliers. We obtain a mean vector and covariance matrix with missing and contaminated data and compare the results of imputation methods using mean square errors. As an application, we use the regression data and examine the effect of missingness on regression model's parameters. We compare the imputed values with real values and explain the results of classical and robust imputation methods.

Keywords: *ER Algorithm, missing data, outliers, robust imputation, sequential imputation.*

1. INTRODUCTION

Statistical methods are used in many research fields. Although the mostly used statistical methods involve numerous assumptions, researchers have been trying to find solutions to the violation of parametric assumptions. Some nonparametric methods can be used to solve assumption problems, but they will also be ineffective when the data matrix is subject to missingness [23].

Lack of measurement values in data may be due to several reasons, such as fasilure of analytical instruments, inability to accomplish an expensive sampling procedure, and unexpected changes in experimental conditions [20], and it is not easy to deal with all such malfunctions. Discarding observations with missing values or outliers may dramatically reduce the number of observations. Instead of deleting, another way is to use the variables' mean or any other location estimator to avoid missing values. On the one hand, deleting or filling out missing values with location

♠Corresponding author, e-mail: onur.toka@hacettepe.edu.tr

estimators may give misleading results when the data matrix contains contamination. Accordingly, classical methods cannot impute well under the presence of outliers. Therefore, researchers have suggested some robust imputation methods to overcome the mentioned issues ([4], [5], [10], [11]).

Afifi and Elashoff [1] briefly summarized the missing data literature. Rubin [18] formalized the mechanism of missing data, and almost all the imputation methods that followed have been proposed using this mechanism. Little and Rubin [14], Schafer [19], and Allison [2] provide a comprehensive overview of methods. Missing methods and missingness have been studied on a wide range [8, 15, 16, 17, 26]. Expectation maximization (EM) algorithm [6] is one of the most well-known method and its robust version is the expectation robust (ER) algorithm [13]. Moreover, the sequential imputation method [22] and its robust version, robust imputation [4], are methods that successfully deal with missing data. Verboven et al. [22] and Branden and Verboven [4] study some important applications using known gene-expression data sets [21]. However, they only considered the imputation methods based on classification and gene data sets. Multivariate statistical methods also require imputation methods; these are discussed extensively in the literature. Thus, we know that if imputation can be performed properly, statistical methodology will benefit from it.

In this study, we provide a practical application to the estimation of mean vector and covariance matrix at different levels of contamination and missing values. In addition, in a simulation study, we compare classical and robust imputation methods by calculating the MSE additively; the results of our simulation study are provided in tables in the subsequent sections. The next section presents a brief overview of imputation methods. The following section estimates the mean vector and covariance matrix and compares them. Furthermore, a regression model is estimated with M-regression, a robust estimation procedure for imputed data, and imputations and residuals are compared for the same data. The final section discusses the results and offers some suggestions for handling missing data.

2. MISSING DATA AND METHODS

2.1. Missing Data Mechanism and Pattern

The literature presents numerous conventional and modern imputation methods, all beginning by explaining and classifying the missingness mechanism. Assume that $Y = (y_{ij})$ denotes a data set and $M = (m_{ij})$ denotes a missing data indicator matrix, such that $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is observed [12]. The missingness mechanism is symbolized by the conditionality of M and Y. Missing completely at random (MCAR) means that the missingness does not relate to the data itself; thus, MCAR can be notated as $f(M|Y) = f(M)$. Missing at random (MAR) means that the missingness relates to an observed part of data, not to a missing part, and can be notated as $f(M|Y) = f(M|Y_{obs})$, when Y_{obs} includes the observed component of Y_{obs}. Note that most of the missing data methods give better results under the missing at random mechanism. Not missing at random (NMAR) means that the missingness relates to both observed and missing parts of data. One can deduce that missing data methods give better results when they correspond to the missingness mechanism. In the following subsections, we explain in more detail the relation between missing data methods and the missingness mechanism. Missingness pattern can be diversified according to missing values. Univariate nonresponse, monotone, general, file matching are major ones for missing data pattern. Univariate nonresponse is obtained when there is single incomplete Y_K. Monotone pattern is usually obtained in longitudinal studies and clinical trials when observations drop out prior to the end of the study for unknown reasons. There are some specific imputation methods such as cold deck [24] and mixed model expectation algorithm [12]. General type pattern can be obtained in numerous data sets and it usually has MCAR and MAR condition. In our analysis, all of missingness and missing data methods are useful for general type missing pattern and missing values have MAR condition to facilitate the comparisons.

2.2. Deletion methods

Missing data methods have been investigated in the literature, and numerous useful methods have been proposed in the course of the last few decades. However, deletion methods are used in some survey or specific study areas involuntarily due to save time. Before using deletion methods, some restrictions should be revised. For efficient results in deletion methods, missingness has to be MCAR and the proportion of the missing part should be as little as possible. Moreover, obtaining a full data matrix through listwise deletion implies exclusion of the entire missing part of data from analysis. Similarly, pairwise deletion allows researchers to analyze data, obtain all the observed parts of variables separately, and combine the results. Previous studies ([1], [14]) clearly show that pairwise deletion almost always works better than listwise deletion methods. Because it deletes less data than listwise deletion methods, pairwise deletion probably gains more information from data. As their definitions indicate, deletion methods do not impute, but just delete the missing parts of data.

Deletion methods are advantageous to overcome the missingness problem, but they could lead to loss of information. The methods can delete some observed and significant values while deleting missing observations. Moreover, they could lead to underestimating scale problems and narrower confidence intervals. In brief, deletion problems only support to deal with missing part of data.

2.3. Imputation methods

To deal with the information loss and undesirable results of deletion methods, researchers are now replacing missing values with substituted values. In the words of Dempster and Rubin [7], the idea of imputation is both seductive and dangerous. It is seductive because it lulls users into the pleasurable belief that the data are complete after all, and it is dangerous because it lumps together situations where the problems are minor and can be legitimately handled in this manner and situations where standard estimators could have substantial bias when applied to real and imputed data. Deciding on which imputation method to use is the most important issue in this regard. This depends on the proportion, mechanism, and type of missing data. Researchers have to know the missing part and structure of the data to decide on the imputation method. This study analyzes the methods and their features and compares them.

2.3.1. Mean and Median Imputation

Mean imputation is an easy approach to a full data matrix. Wilks [25] proposed this method for a data matrix having a small proportion of missing data. However, it should be avoided using the mean or median imputation method without knowing the pattern of missing data or proportion of missingness. Mean imputation can be used to obtain the mean of observed values for every variable and then impute all the missing variables with their own means. Mathematically, if $m_{ij} = 1$ for every variable j, the value of this observation equals the mean of j in the observed column ($y_{ij} = \overline{y_j}$). Other location estimators (median or mode for some likert measurement) can be used for missing values and may be called median or mode imputation. This can be useful for obtaining a full data matrix, but it may corrupt all deviation estimations even for the MCAR condition. As mathematical representations show, if every missing value is equal to the same value (mean or others), the variance or other estimation results may be smaller than expected. Several correction formulas are used to avoid narrower confidence intervals for variance and covariance matrix, but they give similar results as pairwise deletion methods. If the model-based imputation methods give the same results as deletion methods, it makes no sense to use them. Modelling methodology and computer technology enable us to easily utilize better methods.

2.3.2. Model-Based Methods: Expectation Maximization and Expectation Robust Algorithm

Once researchers obtain similar results from both deletion and mean imputation methods, they focus on the information coming from observed data. If the data have some statistically specific distribution, the missing part can be obtained from model-based methods. The most well-known model-based procedure is the EM algorithm. Dempster et al. [6] use this iterative method with the expectation (E-step) and maximization (M-step) part. Its steps consist of replacing missing data with estimated values, estimating the parameters, re-estimating the missing values, assuming the new parameter estimates are correct, re-estimating the parameters, and iterating until convergence [14]. Briefly, assume that Y_{obs} is the observed part and Y_{mis} is the missing part of the data generated from Y_{obs}. θ_t represents all parameters of distribution and $f_{\theta_t}(Y_{obs})$ and $f_{\theta_t}(Y_{mis})$ are the assumed probability distribution at t_{th} iteration. θ_t is the current parameter of θ. Then, EM algorithm computes the complete-data log-likelihood $\ell(\theta_t|Y)$ and obtains $\emptyset(\theta|\theta_t) = \int f_{\theta_{t+1}}(Y_{mis}|Y_{obs}, \theta = \theta_t)\ell(\theta_t|y)dY_{mis}$ for the t_{th} iteration at E-step. Then, it finds θ that maximized $\emptyset$ at M-step. Through iterations until convergence, we finally obtain the missing values.

Assuming parametric restrictions in EM algorithm, outliers and contaminated data influence the estimated parameters. To avoid this problem, Little and Smith [13] proposed the ER algorithm to replace EM algorithm. They changed the M-step to obtain more resistant estimators in contaminated data by adding a weighted estimator based on the Mahalanobis distance and calculated weights using the Hampel bounded influence function [9]. Cheng and Victoria-Feser [5] clarified that ER algorithm imputes the missing part quite well, but with breakdown point 1/(p + 1) approximately, for p is the parameter number. This means that if the breakdown point decreases, the robust properties of ER algorithm will fail. Briefly, at iteration t, $\theta^{(t)} = \left(\mu^{(t)}, \Sigma^{(t)}\right)$ is the current parameter estimation. If observation i contains missing values, $X_{(oi)}$ is the observed array part for X variables and o_i, the number of observations. Note that $\mu_{(oi)}$ and $\Sigma_{(ooi)}$ are the $(o_i \times 1)$ dimensional mean vector and the $(o_i \times o_i)$ dimensional covariance matrix, respectively. The E-step for ER algorithm gives sufficient statistics for mean and covariance with equations (1) and (2):

$$E\left\{\sum_{i=1}^{n} X_{ij}|X_{(oi)}, \theta^{(t)}\right\} = \sum_{i=1}^{n} X_{ij}^{(t)} \qquad j = 1,2,\dots,p \qquad (1)$$

$$E\left\{\sum_{i=1}^{n} X_{ij}X_{ik}|X_{(oi)},\theta^{(t)}\right\} = \sum_{i=1}^{n}\left\{X_{ij}^{(t)}X_{ik}^{(t)} + C_{jki}^{(t)}\right\} \qquad j,k = 1,2,\dots,p \qquad (2)$$

In equations (1) and (2),

$$X_{ij}^{(t)} = \begin{cases} X_{ij}, & X_{ij}\ is\ observed \\ E\left[X_{ij}|X_{(oi)},\theta^{(t)}\right], & X_{ij}\ is\ missing \end{cases}$$

$$C_{jki}^{(t)} = \begin{cases} 0, & X_{ij}\ or\ X_{ik}\ is\ observed \\ cov\left[X_{ij}X_{ik}|X_{(oi)},\theta^{(t)}\right], & X_{ij}\ or\ X_{ik}\ is\ missing \end{cases}$$

We find the imputation values $E\left[X_{ij}|X_{(oi)},\theta^{(t)}\right]$ and correction values $C_{jki}^{(t)}$ by applying the sweep operator to $\Sigma^{(t)}$ [3].

When the weights $w_i = {w(d_i)}/{d_i}$ at R-step can be found from the Mahalanobis distance (d_i) and the Hampel bounded influence function ($w(d_i)$), we obtain the robust estimation of parameters as equations (3) and (4), and then repeat all the steps (iteration) until convergence:

$$\mu_j^{(t+1)} = \sum_{i=1}^{n} w_i X_{ij}^{(t)} \Big/ \sum_{i=1}^{n} w_i \qquad (3)$$

$$\sigma_{jk}^{(t+1)} = \frac{\sum_{i=1}^{n} w_i^2\left(X_{ij}^{(t)} - \mu_j^{(t+1)}\right)\left(X_{ik}^{(t)} - \mu_k^{(t+1)}\right) + C_{jki}^{(t)}}{\sum_{i=1}^{n} w_i^2 - 1} \qquad (4)$$

2.3.3. Sequential and Robust Imputation

While we can use deletion and imputation methods, they have some disadvantages such as infeasibility, computation time problems, and inaccurate solution. To overcome problems, Verboven et al. [22] proposed the sequential imputation (SEQimpute) method. In this method, the covariance matrix and determinant play an important role. $s_{ij} = cov(X_i, X_j)$ is the covariance measure for linear dependency between variables i and j, and we also know that $s_{ii} = cov(X_i, X_i) = var(X_i)$ is the sample variance for variable i. For the multivariate data set, S matrix includes all s_{ij} and s_{ii}. Those who refer to the method assume that a smaller determinant for variance-covariance matrix results in more accurate imputed values. In the algorithm steps, first, the covariance matrix has to be calculated, $s = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T$, and then its determinant D, where $\bar{X}$ is the sample mean. Verboven et al. [17] proposed imputation methods taking advantage of the determinant of the covariance matrix. This method sequentially estimates the missing values, taking one new incomplete observation.

Let us assume that the complete data matrix X_c, the missing values in an observation, $x^* = [x_m^{*T} \quad x_o^{*T}]^T$, are imputed sequentially by minimizing the determinant of the data covariance, $X^* = [X_c^T \quad x^*]^T$. Minimizing the determinant of the complete data matrix part with respect to x_m^* is equivalent to minimizing $D(x^*) = (x^* - \bar{x}_c)^T(cov(X_c))^{-1}(x^* - \bar{x}_c)$, where $\bar{x}_c$ is the row mean estimate. By solving the minimization problem as detailed in [22], x^* is completed and then included in the complete data set. Thus, X_c is replaced in the new complete data part, $X_c = [X_c^T \quad x^*]^T$. Iteration continues until all missing values are imputed.

Branden and Verboven [4] proposed a robust version of the SEQimpute, and called it the robust imputation (ROBimpute) method. All the steps are the same as in the SEQimpute method, but they used the initial robust mean and covariance matrix of the complete data matrix. They proposed using a $(1-\alpha)$% part (α is the contaminated part) of the observed data, but also found the robust estimators useful. The quantity of

contamination (α) can be found with the outlyingness equation, as in [11].

In the next section, we compare all the missing data methods using simulated data and a well-known contaminated regression data. We also discuss the mean vector and covariance matrix estimation for outliers and contamination for imputation methods.

3. SIMULATION STUDY AND APPLICATION

3.1. Simulation study for missing data methods

In our simulation study, we generate a data matrix with 3 variables and 50 observations using multivariate normal distribution $MN(\underline{0}, I)$. The data were contaminated by a proportion of 10% and 20% respectively by data generation from multivariate normal distribution $MN(\underline{5}, I)$. In addition, there was missingness from excluding some values missing randomly for every data set by a proportion of 5% and 10% respectively. We observe the mean vector and covariance matrix randomly for only one simulation design to get comments easily. At the end of generating the simulated data sets, we obtain six different data structures to impute the missing values and estimate the parameters. The imputation methods included mean square errors (MSE) for 500 iterations. Notwithstanding all these, the contaminated and missingness data sets were run for $p = 4$ and $n = 100$; since there were no differences for comments and comparisons, they are not given here.

Table 1 gives the mean vector and covariance matrix estimation for data with no contamination and 5% and 10% missing data at random. The variables' mean of square errors are also given. Because the data show no contamination and least missing proportion, the MSE of the estimations are similar, except for ER algorithm. There is no suggestion for this structure, and all the methods give almost similar results. Deletion methods lead to loss of some parameter information because of deletion of the observation list or some observed values. Mean and median imputation give the lowest MSE compared to the other methods. Mathematically, mean vector estimation for pairwise deletion and mean imputation has to be the same. On the contrary, setting all missing values with the mean for each variable leads to underestimation of the covariance matrix compared to pairwise deletion methods.

As the missingness part increases, it becomes clearer that the SEQimpute and ROBimpute methods give the same results. This is as expected, because the proportion of contamination is zero. It is clear that when the data values are MAR and no contamination exists, the results are the same. As the missingness part increases, model-based methods give smaller MSE than single value imputation and deletion methods. The only unexpected case in Table 1 is that of the ER algorithm, which gives misleading estimation with no contamination.

Table 1. Results for No Contamination and 5%, 10% Missing Data

	No contamination 5% missing					No contamination 10% missing				
Methods	$\bar{X}^T$		$cov\,(X)$		MSE	$\bar{X}^T$		$cov\,(X)$		MSE
	0,0682	0,7553	-0,1530	-0,1350	**0,0076**	0,0736	1,4893	0,0872	-0,1890	0,0094
Listwise	-0,0104	-0,1530	1,3802	-0,0098		-0,1543	0,0872	0,9174	0,0645	
Deletion	0,0511	-0,1350	-0,0098	1,0288		0,0867	-0,1890	0,0645	1,4288	
	-0,0369	0,7785	-0,1598	-0,1500	**0,0069**	0,1304	1,4626	0,1445	-0,1508	0,0082
Pairwise	-0,0073	-0,1598	1,4069	-0,0098		-0,1202	0,1445	0,8834	0,0399	
Deletion	0,0440	-0,1500	-0,0098	0,9690		0,0104	-0,1508	0,0399	1,4867	
	-0,0369	0,7785	-0,1468	-0,1408	**0,0069**	0,1304	1,3432	0,1211	-0,1272	0,0082
Mean	-0,0073	-0,1468	1,2920	-0,0084		-0,1202	0,1211	0,7752	0,0325	
Imputation	0,0440	-0,1408	-0,0084	0,9096		0,0104	-0,1272	0,0325	1,3350	
	-0,0369	0,7785	-0,1414	-0,1389	**0,0069**	0,1376	1,3438	0,1184	-0,1300	**0,0079**
Median	-0,0149	-0,1414	1,2927	-0,0080		-0,1344	0,1184	0,7767	0,0387	
Imputation	0,0407	-0,1389	-0,0080	0,9098		0,0215	-0,1300	0,0387	1,3361	
EM	-0,0369	0,7630	-0,1651	-0,1423	**0,0070**	0,1376	1,4317	0,1363	-0,1553	**0,0078**

Algorithm	0,0063	-0,1651	1,3781	-0,0036		-0,1187	0,1363	0,8633	0,0355	
	0,0511	-0,1423	-0,0036	0,9475		0,0096	-0,1553	0,0355	1,4526	
	0,3421	0,6429	-0,0979	-0,1345	0,0141	0,1706	0,5681	-0,0359	0,1568	0,0151
ER	-0,1160	-0,0979	1,2569	-0,1160		0,0631	-0,0359	0,9493	0,1709	
Algorithm	-0,1545	-0,1345	-0,1160	0,8016		0,1330	0,1568	0,1709	0,7724	
	-0,0369	0,7785	-0,1683	-0,1454	**0,0070**	0,1353	1,3460	0,1303	-0,1558	**0,0078**
SEQimput	0,0066	-0,1683	1,2967	-0,0041		-0,1215	0,1303	0,7759	0,0337	
	0,0515	-0,1454	-0,0041	0,9106		0,0123	-0,1558	0,0337	1,3358	
	0,0150	0,9349	-0,2803	-0,3502	**0,0070**	0,0628	1,2112	0,3318	-0,4252	**0,0078**
ROBimput	0,0864	-0,2803	1,3079	0,3310		-0,1696	0,3318	0,9596	0,0857	
	0,0845	-0,3502	0,3310	0,7436		-0,0284	-0,4252	0,0857	1,7664	

Table 2 gives the mean vector and covariance matrix estimation for data with 10% contamination and 5% and 10% missing data at random. The table shows that robust properties made the MSE smaller and contamination spoiled classic estimation. ER algorithm and robust imputation give the lowest MSE, and all the classical estimations of the covariance matrix are distorted.

Table 2. Results for 10% Contamination and 5%, 10% Missing Data

	10% contamination 5% missing					10% contamination 10% missing				
Methods	$\bar{X}^T$		$cov\,(X)$		MSE	$\bar{X}^T$		$cov\,(X)$		MSE
	0,4548	3,3149	1,7943	1,9212		0,5697	3,5793	2,6675	2,5601	
Listwise	0,6201	1,7943	2,5380	1,8933	0,0953	0,4214	3,5793	2,6675	2,5601	0,1071
Deletion	0,4221	1,9212	1,8933	2,7690		0,4601	3,5793	2,6675	2,5601	
	0,3994	3,1836	1,7232	1,9072		0,5188	3,5579	3,0910	2,3827	
Pairwise	0,5651	1,7232	2,4415	1,8933	0,0915	0,4559	3,0910	4,2802	2,4736	0,0942
Deletion	0,4081	1,9072	1,8933	2,7165		0,3987	2,3827	2,4736	2,8436	
	0,3994	3,1186	1,6528	1,7515		0,5188	3,1223	2,4643	1,9948	
Mean	0,5651	1,6528	2,3418	1,7008	0,0915	0,4559	2,4643	3,7561	2,0646	0,0942
Imputation	0,4081	1,7515	1,7008	2,4947		0,3987	1,9948	2,0646	2,6695	
	0,3896	3,1235	1,6757	1,7603		0,4569	3,1510	2,5410	1,9739	
Median	0,5470	1,6757	2,3499	1,7329	0,0854	0,3856	2,5410	3,7930	2,0858	0,0813
Imputation	0,3719	1,7603	1,7329	2,5100		0,3826	1,9739	2,0858	2,6736	
	0,3994	3,1186	1,6863	1,7890		0,5181	3,3034	2,6037	2,3671	
EM	0,5483	1,6863	2,3772	1,7553	0,0906	0,3803	2,6037	3,9996	2,5861	0,0920
Algorithm	0,3767	1,7890	1,7553	2,5993		0,4445	2,3671	2,5861	3,0035	
	0,3662	**1,7514**	**1,3212**	**1,1103**		**0,4423**	**0,7536**	**0,1316**	**0,5053**	
ER	**0,0387**	**1,3212**	**1,9250**	**0,8673**	**0,0648**	**0,3580**	**0,1316**	**1,3148**	**0,0750**	**0,0768**
Algorithm	**0,2667**	**1,1103**	**0,8673**	**2,1332**		**-0,1703**	**0,5053**	**0,0750**	**1,4289**	
SEQimpute	0,3994	3,1186	1,6843	1,7891	0,0904	0,5214	3,1716	2,5911	2,3777	0,0920
	0,5493	1,6843	2,3538	1,7551		0,3802	2,5911	3,8356	2,5984	

	0,3762	1,7891	1,7551	2,5294		0,4455	2,3777	2,5984	2,9797	
	-0,1655	**1,1658**	**0,2513**	**-0,0223**	**0,0839**	**-0,0448**	**0,9167**	**0,1806**	**0,1111**	**0,0784**
ROBimput	**0,2155**	**0,2513**	**1,2290**	**0,4414**		**-0,2043**	**0,1806**	**0,8560**	**0,2854**	
	-0,0609	**-0,0223**	**0,4414**	**1,1443**		**-0,0584**	**0,1111**	**0,2854**	**1,4669**	

Table 3 gives the mean vector and covariance matrix estimation for data with 20% contamination and 5% and 10% missing data at random. The table shows that robust properties made the MSE smaller and contamination spoiled classic estimation. ER algorithm and robust imputation give the lowest MSE, and all the classical estimations of the covariance matrix are distorted.

As in Table 2, median imputation has smaller MSE for mean estimation in non-model-based methods, but its covariance matrix is more affected than robust procedures.

Table 3. Results for 20% Contamination and 5%, 10% Missing Data

	20% contamination 5% missing					20% contamination 10% missing				
Methods	$\bar{X}^T$		$cov\ (X)$		MSE	$\bar{X}^T$		$cov(X)$		MSE
	1,2652	4,8934	4,1718	3,8036	0,3422	0,8484	4,6377	3,5489	3,5644	
Listwise	0,7238	4,1718	4,6120	3,6204		0,7584	3,5489	4,8294	3,8003	0,3580
Deletion	0,8683	3,8036	3,6204	4,2985		0,8826	3,5644	3,8003	4,4019	
	1,2432	4,6703	3,9735	3,7218	0,3411	0,9737	4,9222	3,4098	4,1663	
Pairwise	0,8078	3,9735	4,7575	3,9152		0,7485	3,4098	4,9102	4,0449	0,3450
Deletion	0,9241	3,7218	3,9152	4,4208		1,0320	4,1663	4,0449	5,0389	
	1,2432	4,4797	3,5716	3,4141	0,3411	0,9737	4,4200	2,7135	3,4867	
Mean	0,8078	3,5716	4,4662	3,5158		0,7485	2,7135	4,3090	3,3839	0,3450
Imputation	0,9241	3,4141	3,5158	4,2403		1,0320	3,4867	3,3839	4,6276	
	1,2083	4,5095	3,4820	3,3887	0,3199	0,9005	4,4692	2,6706	3,5296	
Median	0,7804	3,4820	4,4783	3,5242		0,6882	2,6706	4,3362	3,3315	0,3020
Imputation	0,9047	3,3887	3,5242	4,2495		0,9719	3,5296	3,3315	4,6699	
	1,3161	4,8365	4,0971	3,7707	0,3398	0,9602	4,9254	3,9927	3,9837	
	0,8264	4,0971	4,5743	3,6075		0,8601	3,9927	5,2090	4,2402	0,3510
EM Algorithr	0,8918	3,7707	3,6075	4,2282		1,0010	3,9837	4,2402	4,7610	
	0,6092	**1,3154**	**0,8947**	**1,0672**	**0,3101**	**0,5545**	**3,2447**	**2,1553**	**2,4214**	
	0,2538	**0,8947**	**1,5168**	**0,8505**		**0,6075**	**2,1553**	**2,6740**	**2,2882**	**0,2740**
ER Algorithm	**0,3959**	**1,0672**	**0,8505**	**1,9264**		**0,9281**	**2,4214**	**2,2882**	**3,2493**	
	1,3162	4,9056	4,1822	3,8511	0,3396	0,9583	4,7947	3,9920	3,9772	
SEQimpute	0,8250	4,1822	4,6019	3,6777		0,8624	3,9920	5,0749	4,2269	0,3500
	0,8925	3,8511	3,6777	4,2673		0,9988	3,9772	4,2269	4,6885	
	0,3752	**0,6287**	**-0,0188**	**0,0812**	**0,3180**	**0,0791**	**1,2118**	**0,1783**	**0,1910**	
ROBimpute	**-0,0932**	**-0,0188**	**0,9740**	**0,2051**		**-0,1631**	**0,1783**	**1,5566**	**0,4461**	**0,2970**
	-0,0498	**0,0812**	**0,2051**	**1,4784**		**0,1026**	**0,1910**	**0,4461**	**1,0527**	

For all data structures, Figure I gives the MSE for imputation methods for all data structures. For contamination, we find that ER algorithm and robust imputation are comparatively better methods.

3.2. Application of imputation methods

To apply imputation methods on well-known data, we use data from [10]. The data have 75 observations, of which the first 14 are contaminated. Because of missingness of data, we deleted 22 values randomly. All the missingness parts are imputed with mean imputation, sequential imputation, ER algorithm, and robust imputation. The results are listed in Table 4. In particular, one of the missing values is from the contaminated part. Because this value is an outlier, mean imputation methods cannot approximate the real value. The other methods give better results.

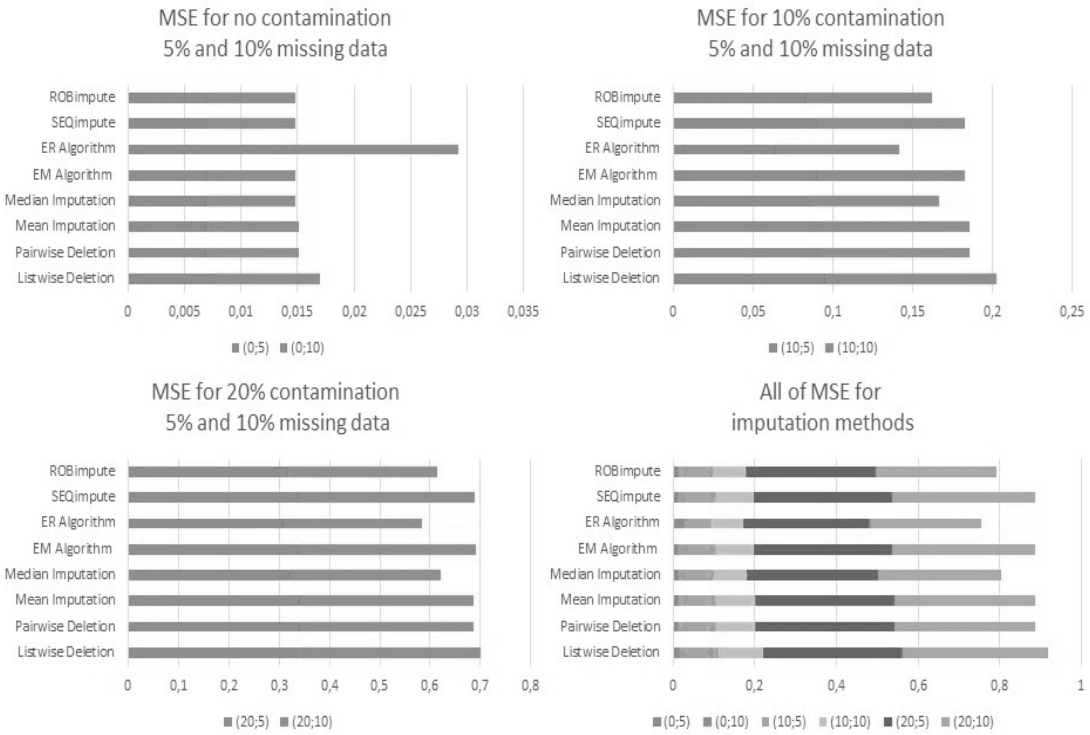

Figure 1. All of MSE Results for 0%, 10%, 20% Contamination and 5%, 10% Missing Data

For easier comparison of results, we give the mean of imputation error for every imputation method in Table 5. Robust imputation and ER algorithm give similar results. Since mean imputation failed to impute the outlier and missing values, it means the imputation error increased more than the others.

Table 4. Imputation Results for Hakwins et al. (1984)'s Data

Variable	Number of observation	Real Value	Mean Imputation	Sequential Imputation	ER Algorithm	Robust Imputation
x_3	**9**	31,00	7,32	26,15	29,37	29,43
x_1	**15**	3,40	3,30	1,90	1,57	1,42
x_2	**18**	1,60	5,99	1,91	2,16	1,98
x_1	**22**	0,40	3,30	3,39	2,10	2,11
x_2	**22**	3,20	5,99	7,16	3,84	3,83
y	**22**	0,30	1,32	-1,38	0,59	0,70
x_3	**32**	0,30	7,32	3,30	1,88	1,75
x_1	**41**	3,40	3,30	2,84	1,70	1,60

x_2	**41**	1,60	5,99	1,19	2,51	2,41
x_2	**46**	0,50	5,99	2,53	2,68	2,65
y	**46**	-0,40	1,32	-1,17	0,30	0,38
y	**49**	0,90	1,32	0,26	0,31	0,42
x_3	**50**	2,90	7,32	5,51	2,73	2,36
x_2	**51**	1,50	5,99	1,84	1,50	1,50
x_2	**52**	0,60	5,99	-0,64	1,78	1,59
x_1	**54**	1,10	3,30	5,85	2,10	2,07
x_3	**54**	0,30	7,32	8,09	3,73	3,66
x_1	**58**	2,40	3,30	1,81	1,64	1,66
x_1	**59**	1,60	3,30	1,45	1,60	1,54
x_1	**64**	2,80	3,30	0,05	1,64	1,56
x_3	**66**	0,80	7,32	0,03	0,86	0,71
x_2	**74**	2,20	5,99	3,03	1,35	1,27

Table 5. Mean of Imputation Error for Methods

Imputation Methods	**Mean of Imputation Error**
Mean Imputation	40,13
Sequential Imputation	7,68
ER Algorithm	1,74
Robust Imputation	1,72

Figure 2 depicts the differences between imputations. It shows that robust imputations and ER algorithm give the best imputation for missing values in contaminated data.

4. CONCLUSION

Researchers often encounter the missingness problem. This has been overcome through deletion or imputation methods. Missing data methods require several assumptions, and so the other statistical methods and assumptions should not be ignored. Classical methods can guarantee the results in case of assumptions. Therefore, some cases such as outliers or contaminated data corrupt the estimation of classical imputation methods and therefore parameter estimation. As can be seen from simulation studies and their applications using well-known data, robust imputation methods can cope well with outliers to impute missing values and estimate parameters, as the ER algorithm mean and covariance estimation shows. However, researchers should know the percent of outliers or contaminated part in the data for ER algorithm. After imputing values with robust imputation, robust estimators can be used to obtain more efficient estimation results compared to the classical methods, as shown in our application part.

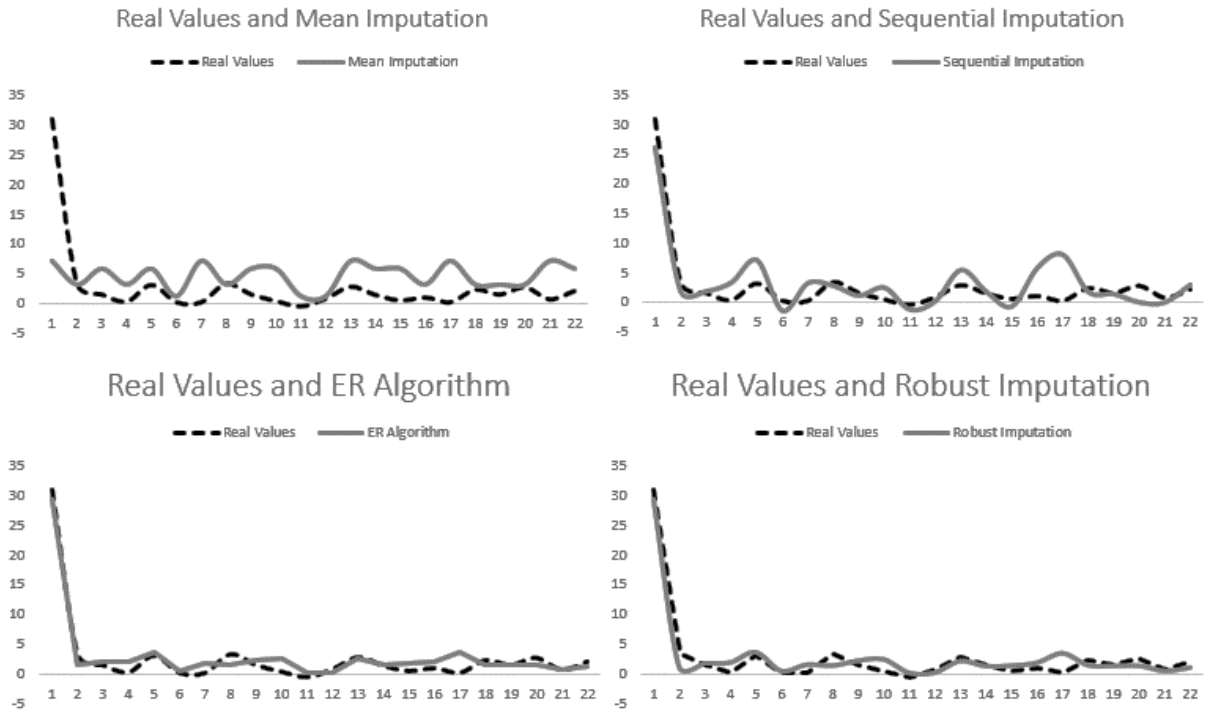

Figure 2. Differences between real values and imputation methods

Missingness is not a major problem if you know how to handle it. Numerous new suggestions show how to handle special data such as biostatistics, data-microarrays, and electrical data. In this study, we presented and compared some missing methods that can be used for continuous data types. The results showed that ER algorithm and robust imputation can solve the missing data problem in case of outliers or contamination.

CONFLICT OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

[1] Afifi, A. A. and Elashoff, R. M., "Missing observations in multivariate statistics I. Review of the literature, Journal of the American Statistical Association", 61:595-605, (1966).

[2] Allison, P. D. "Missing data: Quantitative applications in the social sciences. British Journal of Mathematical and Statistical Psychology", 55(1): 193-196, (2002).

[3] Beale, E. M. L., Little, R. J. A. "Missing values in multivariate analysis, Journal of the Royal Statistical Society, Series B", 37:129-145, (1975).

[4] Branden, K., Verboven, V. S., "Robust data imputation, Computational Biology and Chemistry", 33(1): 7-13, (2009).

[5] Cheng, T. S., Victoria-Feser, M. P. "High-breakdown estimation of multivariate mean and covariance with missing observations", British J. Math. Statist. Psych., 5: 317–335, (2002).

[6] Dempster, A. P., Laird, N. M., Rubin, D. B., "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, Series B, 39: 1-38, (1977).

[7] Dempster, A. P., Rubin, D. B. 1983, "Introduction of incomplete data in sample surveys (Volume 2)" Theory and Bibliography (W. G. Madow, I. Olkin, D.B. Rubin eds.)", 3-10, New York.

[8] Graham, J.W., Missing Data: Analysis and Design, Springer New York, 324 p., (2014).

[9] Hampel, F. R. "The influence curve and its role in robust estimation", The Annals of Statistics, 69: 383–393, (1974).

[10] Hawkins, D.M., Bradu, D. and Kass, G.V. "Location of several outliers in multiple regression data using elemental sets". Technometrics, 26: 197–208. (1984).

[11] Hubert, M., Rousseeuw, P. J. and Vanden Branden, K., "ROBPCA: a new approach to robust principal component analysis", Technometrics, 47(1): 64-79, (2005).

[12] Ibrahim, J.G. and Molenberghs, G., "Missing Data Methods in Longitudinal Studies: A Review, Test (Madrid, Spain)", 18.1:1–43, (2009).

[13] Little, R. J. A., Smith, P. J., "Editing and imputing for quantitative survey data", Journal of the American Statistical Association 82:58-68, (1987).

[14] Little, R. J. A., Rubin, D. B., Statistical Analysis with Missing Data (2nd ed.), Hoboken, N. Jersey, Wiley, (2002).

[15] Lynch, S.M. and Bron, J.S., "Handling Missing Data in Social Research", Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences, (2015).

[16] O'Kelly, M. and Ratitch, B., "Clinical Trials with Missing Data: A Guide for Practitioners", John Wiley & Sons, (2014).

[17] Raghunathan, T., "Missing Data Analysis in Practice", Chapman & Hall CRC Interdisciplinary Statistics, (2015).

[18] Rubin, D. B. "Inference and missing data", Biometrika, 63:581–592, (1976).

[19] Schafer, J. L., "Analysis of incomplete multivariate data", Boca Raton, FL: Chapman & Hall, (1997).

[20] Stanimirova, I. and Walczak, W., "Classification of data with missing elements and outliers", Talanta, 76, 602-609, (2008).

[21] Toka, O., Kayıp Veri Durumunda Sağlam Kestirim, H.Ü. Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, Ankara, (2012).

[22] Verboven, S., Branden, K.V. and Goos, P. "Sequential imputation for missing values", Computational Biology and Chemistry, 31:320-327, (2007).

[23] Wang, D. and Chen, S. X., "Nonparametric imputation of missing values for estimating equation based inference", Statistics Preprints, Paper 41, (2005).

[24] Wang, J., Data Mining: Opportunities and Challenges, Idea Group Inc (IGI), (2003).

[25] Wilks, S. S., "Moments and distributions of estimates of population parameters from fragmentary samples", The Annals of Mathematical Statistics, 3:163–195, (1932).

[26] Zhou, X., Zhou, H. C., Lui, D. and Ding, X. Applied Missing Data Analysis in the Health Sciences, John Wiley & Sons, (2014).