Missing Values Imputation Using Principal Component Analysis Methods

Rhoda Josephina Domebale Moh

Department of Mathematical Sciences
Montana State University

May 9, 2024

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Rhoda Josephina Domebale Moh

This writing project has been read by the writing project advisor and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

Date	Dr. Samidha Shetty Writing Project Advisor
Date	Dr. Katharine Banner Writing Projects Coordinator

Abstract

Missing values are a common phenomenon encountered in datasets, posing challenges to data analysis. Thus, it becomes important to employ effective methods for imputing missing values to reduce potential bias in data analysis. Principal Component Analysis (PCA) is a well-known technique for reducing data dimensionality. However, there have been instances where PCA has been used for imputing missing data. In this project, we explore three different PCA-based methods namely Singular Value Decomposition PCA (SVDPCA), Probabilistic PCA (PPCA) and Local Least Squares PCA (LLSPCA). These methods are applied to a dataset with two types of missingness: missing completely at random (MCAR) and missing at random (MAR). The performance of these methods will be discussed under the MCAR and MAR assumptions, for different percentages of missing values in the data.

Contents

1 Introduction

Missing values occur when there is no value recorded for a variable in a data set. There are various reasons why data can have missing values. Some of these reasons include human error, data entry errors, respondents failing to provide information during data collection process, malfunctioning of equipment among others. They are a common phenomenon in real life data and can pose several challenges to data analysis, such as bias in results, loss of statistical power and reduction of representativeness of the data (Hyun (2013)). Therefore, it is crucial to employ effective techniques to impute missing values prior to data analysis.

Prior to missing value imputation, it is important to investigate the underlying mechanism that generated the missing values. The types of missing values were first described by Rubin (1976) and they fall into three categories: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR).

Missing Completely at Random (MCAR) occurs when the probability of missing values is not related to the observed data. Thus, the missingness is entirely random and independent of observed and unobserved data. Missing at Random (MAR) occurs when the probability of missing values depends on the observed data. When missing values are not categorized as MCAR or MAR, then they are put into the category of Missing Not at Random (MNAR), which is considered the most difficult mechanism to work with. For MNAR, the missingness can depend on unobserved variables. The scope of this project covers MCAR and MAR cases.

When addressing missing values, there are several approaches one could employ. Common approaches used include deletion and imputation. Some deletion methods involve removing rows that have missing values from the dataset. However, this method leads to loss of data, which can subsequently lead to loss of power and biased estimates. Some common imputation methods also include mean and median imputation, where missing values within columns of data are replaced with the

respective column mean and median values. A potential problem here as well is that these values are then treated as the true values in the analysis, which can also introduce some bias in the results.

More robust methods for imputation such as mutliple imputation and Principal Component Analysis (PCA) based methods have been developed to address the problems of biased estimates and missing data more effectively. In multiple imputation, proposed by Rubin (1996), missing values are replaced by multiple plausible values based on the distributions of the variables in the data. The multiple imputed datasets are then analyzed separately, and the results are combined to provide more accurate estimates. The estimates from multiple imputation have been found to result in valid statistical inference (Li et al. (2015)). However, because multiple datasets need to be generated for the multiple imputation procedures, this method can be computationally intensive for high-dimensional data (Nguyen et al. (2023)). Thus, PCA-based methods become useful in this context, as PCA reduces data dimensionality while capturing important relationships among variables in the data.

The main objective of this project is to explore three different PCA based imputation methods namely Singular Value Decomposition, Local Least Squares and Probabilistic PCA methods. In Section 2, we introduce the concept of Principal Component Analysis. Then, in Section 3 we explore the three PCA-based methods for imputation of missing data. Finally, in Section 4, these methods will be implemented on a data set and their performance will be compared.

2 Principal Component Analysis (PCA) Overview

Principal Component Analysis (PCA) is a commonly used method for dimension reduction. First introduced by Karl Pearson in 1901 (Pearson (1901)), the goal of

PCA is to reduce dimensionality of a multivariate dataset, while accounting for as much variation in the original dataset as possible (Everitt and Hothorn (2011)). A new set of variables called the principal components are formed, which are linear combinations of the original variables in the dataset. These principal components are uncorrelated variables. Additionally, they are ordered, such that the first few principal components account for the most variation in the original variables (Everitt and Hothorn (2011)).

Suppose there exist a set of variables in a dataset $\boldsymbol{x}^T = (x_1, ..., x_q)$, then a set of uncorrelated variables $\boldsymbol{y}^T = (y_1, ..., y_q)$ are formed which are linear combinations of the original variables, as described below:

$$y_1 = a_{11}x_1 + \cdots + a_{1q}x_q$$

$$\vdots$$

$$y_q = a_{q1}x_1 + \cdots + a_{qq}x_q$$

$\boldsymbol{y}^T$ is ordered in decreasing order of variance such that $Var(y_1) > Var(y_2) > \cdots > Var(y_q)$. For y_1, choose values of $\boldsymbol{a}_1^T = (a_{11}, a_{12}, \cdots, a_{1q})$ that maximizes $Var(y_1)$ such that $\sum_{j=1}^{n} a_{ij}^2 = 1$. For y_2, choose values of $\boldsymbol{a}_2^T = (a_{21}, a_{22}, \cdots, a_{2q})$ that maximizes $Var(y_2)$ such that

1. $\sum_{j=1}^{q} a_{ij}^2 = 1 \leftrightarrow \boldsymbol{a}_j^T \boldsymbol{a}_j = 1$
2. $\boldsymbol{a}_2^T \boldsymbol{a}_1 = 0$

The second condition implies that y_2 is uncorrelated with y_1. In general, for the k^{th} principal component y_k, maximize $Var(y_k)$ such that

1. $\boldsymbol{a}_k^T \boldsymbol{a}_k = 1$
2. $\boldsymbol{a}_k^T \boldsymbol{a}_j = 0$ for all $j < k$

After the principal components have been found, the subsequent important step is determining the optimal number of primary components to use. There is no definitive rule for determining the number of principal components, however, Jolliffe (2002) suggests keeping enough principal components that account for between 70% and 90% of the total variation of the variables.

In the next section, we discuss the three PCA-based methods implemented in this project.

3 Imputation Methods

For this project, three PCA-based imputation methods were implemented, namely Singular Value Decomposition PCA (SVDPCA), Local Least Squares PCA (LLSPCA) and Probabilistic PCA (PPCA) methods. We briefly discuss these methods below. All methods were implemented using the `pcaMethods` (Stacklies et al. (2007)) package in `R` (R Core Team (2023)).

3.1 Singular Value Decomposition PCA (SVDPCA)

In this section, we describe the Singular Value Decomposition PCA (SVDPCA) method as initially proposed by Troyanskaya et al.. The motivation behind this method stems from the need to fill in missing values in gene expression microarray experiments, which often result in data sets with many missing expression values.
This method uses singular value decomposition to obtain principal components from the data matrix which are then used to impute missing values. In singular value decomposition (SVD), an $m \times n$ data matrix $\boldsymbol{X}$ is decomposed into a product of three matrices as follows

$$\boldsymbol{X}_{m\times n} = \boldsymbol{U}_{m\times m}\boldsymbol{\Sigma}_{m\times n}\boldsymbol{V}^{\boldsymbol{T}}_{n\times n} \tag{1}$$

where $\mathbf{U}$ is an $m \times m$ orthogonal matrix, $\mathbf{\Sigma}$ is an $m \times n$ diagonal matrix with real non-negative numbers on the diagonal, and $\mathbf{V}$ is an $n \times n$ orthogonal matrix. $\Sigma_{m \times n} = diag(\sigma_1, \sigma_2, \cdots, \sigma_n)$ where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$.

The columns of $\boldsymbol{V}$ are identical to principal components (PCs) obtained from a classical PCA procedure. They are sorted according to importance based on the corresponding eigenvalues in $\mathbf{\Sigma}$. Then, the k most significant principal components are identified and selected. Troyanskaya et al. (2001) state that exact number of k components should be determined empirically. For our dataset, we run a classical PCA and selected the principal components that account for at least 85% of the variation in the data. After selecting k, the next step is to impute the missing values as linear combinations of the k selected PCs, and we describe this process in the next paragraphs.

For each missing value $\alpha_{i,j}$ in row $i = 1, \cdots, m$ and column $j = 1, \cdots, n$, replace $\alpha_{i,j}$ in $\boldsymbol{X}$ with $\bar{\alpha} = \frac{1}{n} \sum_{j=1}^{n} x_{i,j}$ to obtain $\boldsymbol{X'}$, where $x_{i,j}$ represents the data entries in each i^{th} row across j columns. SVD is then performed on $\boldsymbol{X'}$ using 1.

The next step is to choose k significant PCs from $\boldsymbol{V}^T$. For each $\alpha_{i,j}$ in row i, solve $X_{i,-j} = \beta_0 + \sum_{m=1}^{k} \beta_m Y_{m(-j)}$, where $X_{i,\cdot}$ is the i^{th} row with all elements except the missing value $\alpha_{i,j}$ and $Y_{m(-j)}$ is the m^{th} principal component selected from $\boldsymbol{V}^T$ except the j^{th} elements, for $m \in (1, \cdots k)$. Using the estimated $\hat{\beta}$ and the j^{th} elements of the principal components $Y_{m(j)}$, $\alpha_{i,j}$ is reconstructed as $\alpha_{i,j} = \hat{\beta}_0 + \sum_{m=1}^{k} \hat{\beta}_m Y_{m(j)}$. In other words, the missing value $\alpha_{i,j}$ is estimated by regressing $X_{i,-j}$ against the k PCs and then using the coefficients of the regression to reconstruct $\alpha_{i,j}$ as a linear combination of the k PCs. Once all $\alpha_{i,j}$ in $\boldsymbol{X'}$ are imputed, the process is then repeated on the new matrix until total change in matrix falls below a determined threshold of 0.01.

3.2 Local Least Squares PCA (LLSPCA)

Similar to the SVDPCA method, the Local Least Squares PCA method was also developed to impute missing values in gene microarray experiments. Initially proposed by Kim et al. (2004), the local least squares imputation is based on least squares formulation, which exploits local similarity structures in the data.

The LLSPCA method imputes missing values of a given row in the data as linear combinations of k rows with similar features. These k rows are selected using L_2-norm or Pearson's correlation coefficients. In this project, we will describe the Pearson's correlation coefficient method by Kim et al. (2004) , as implemented in the R package `pcaMethods` (Stacklies et al. (2007)).

The first step is selecting the k nearest neighbors. For simplicity in the algorithm, Kim et al. (2004) describes the missing value estimation by assuming a missing value in the first position of the first row in the data. This missing value is denoted as $\boldsymbol{g}_1(1) = \alpha$ where the first row is denoted as $\boldsymbol{g}_1$. The k-nearest neighbor row vectors for $\boldsymbol{g}_1$ are also defined as

$$\boldsymbol{g}_{s_i}^T \in \mathbb{R}^{1\times n},\ 1 \leq i \leq k$$

where s_i is an index for the selected k-nearest neighbor vectors. For $\boldsymbol{g}_1$, the Pearson correlation coefficient is between two vectors is found as

$$r_{ij} = \frac{1}{n-1}\sum_{k=2}^{n}\big(\frac{g_{1k}-\bar{g}_1}{\sigma_1}\big)\big(\frac{g_{jk}-\bar{g}_j}{\sigma_j}\big) \tag{2}$$

where $\bar{g}_1$ and $\bar{g}_j$ represent the mean values for the row vectors $\boldsymbol{g}_1^T$ and $\boldsymbol{g}_j^T$ respectively, and σ_1 and σ_j represent the standard deviation of these respective row vectors. The columns that correspond to the missing value column are excluded in calculating r_{ij}. Then the k rows with the largest Pearson correlation coefficients in magnitude are selected.

The following are formed from the k-nearest neighbor rows: matrix $A \in \mathbb{R}^{k\times(n-1)}$,

vectors $\boldsymbol{b} \in \mathbb{R}^{k\times 1}$ and $w \in \mathbb{R}^{(n-1)\times 1}$. The matrix $\boldsymbol{A}$ is formed from the k-nearest neighbor row vectors with the first elements excluded, $\boldsymbol{b}$ is formed from the first elements of the k-nearest neighbor row vectors and $\boldsymbol{w}$ is formed from the first row vector $\boldsymbol{g}_1$ with the missing value. For example, for some $N \times 4$ matrix with a missing value in the first element of row 1, suppose k vectors are selected, A, $\boldsymbol{b}$ and $\boldsymbol{w}$ are formed as follows:

$$\begin{pmatrix} \alpha & \boldsymbol{w^T} \\ \boldsymbol{b} & A \end{pmatrix} = \begin{pmatrix} \alpha & w_1 & w_2 & w_3 \\ b_1 & A_{1,1} & A_{1,2} & A_{1,3} \\ \vdots & \vdots & \vdots & \vdots \\ b_k & A_{k,1} & A_{k,2} & A_{k,3} \end{pmatrix}$$

Then the least squares problem can be formulated as follows:

$$\min_x ||A^T \boldsymbol{x} - \boldsymbol{w}||_2 \tag{3}$$

The missing value α is then estimated using the following linear combination

$$\alpha = \boldsymbol{b}^T \boldsymbol{x} = \boldsymbol{b}^T (A^T)^\dagger \boldsymbol{w} \tag{4}$$

where $(A^T)^\dagger$ is the pseudoinverse of A^T and $\boldsymbol{x}$ is a vector of the coefficients of the linear combination obtained from 3.

3.3 Probabilistic PCA (PPCA)

Probabilistic PCA method was introduced by Tipping and Bishop (1999), which extends the classical PCA method to a probabilistic model. PPCA assumes the data can be generated by projecting latent variables into a high-dimensional space. The latent variables are estimated using Maximum Likelihood Estimation (MLE) and an Expectation-Maximization (EM) algorithm. The PPCA method, as derived by Tipping and Bishop (1999) is described below.

For a d-dimensional observation vector $\boldsymbol{t}$ which relates to a corresponding q-dimensional vector of latent variables $\boldsymbol{x}$, we have the model

$$\boldsymbol{t} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

where $\mathbf{W}$ is a $d \times q$ matrix which relates the two variables x and t, $\boldsymbol{x} \sim N(0, \boldsymbol{I_q})$ where $\boldsymbol{I_q}$ is a q-dimensional identity matrix, $\boldsymbol{\mu}$ is a d-dimensional mean vector which permits the model to have non-zero mean, $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I_q})$.

This implies that $t|x$ also follows a Gaussian distribution as follows

$$\boldsymbol{t}|\boldsymbol{x} \sim N(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I_q})$$

The marginal distribution of $\boldsymbol{t}$ also follows a Gaussian distribution, where $\boldsymbol{t} \sim N(\boldsymbol{\mu}, \boldsymbol{C})$, and $\boldsymbol{C} = \sigma^2 \boldsymbol{I} + \boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}}$. The log-likelihood of $\boldsymbol{t}$ is given as

$$\mathcal{L} = \frac{-N}{2} d ln(2\pi) + ln|\boldsymbol{C}| + tr(\boldsymbol{C}^{-\mathbf{1}}\boldsymbol{S})$$

where

$$S = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{t_n} - \boldsymbol{\mu})(\boldsymbol{t_n} - \boldsymbol{\mu})^T$$

S is the sample covariance matrix of observed data $\{\boldsymbol{t_n}\}$, where $n \in \{1, \cdots, N\}$. The MLE of $\boldsymbol{\mu}$ is the mean of the observed data. MLEs for $\boldsymbol{W}$ and $\boldsymbol{\sigma^2}$ can be estimated explicitly or by using an EM algorithm specified by Tipping and Bishop (1999). The `pcaMethods` (Stacklies et al. (2007)) package implements the EM algorithm. Once $\boldsymbol{W}$ and $\boldsymbol{\sigma^2}$ are estimated, the conditional distribution of latent variables $\boldsymbol{x}$ given observed data $\boldsymbol{t}$ is given as

$$\boldsymbol{x}|\boldsymbol{t} = N(\boldsymbol{M}^{-\mathbf{1}}\boldsymbol{W}^{\boldsymbol{T}}(\boldsymbol{t} - \boldsymbol{\mu}), \sigma^2 \boldsymbol{M}^{-\mathbf{1}})$$

where $\boldsymbol{M} = \boldsymbol{W^T W} + \sigma^2 \boldsymbol{M^{-1}}$. Missing values can then be estimated using the values of $\hat{\boldsymbol{x}}, \hat{\boldsymbol{W}}$ and $\hat{\sigma^2}$.

4 Results

4.1 Data

Data on food prices for nutrition was obtained from the World Bank Open Data website . We subset the full dataset to 27 quantitative variables, with no missing data, describing the cost and availability of diets and food groups for 174 countries in 2017.

Using `missMethods` package (Rockel (2022)) in R, missing values were introduced to this data under Missing Completely at Random (MCAR) and Missing at Random (MAR) assumptions across different percentages of missingness (10%, 25%, 40%, 55%). The methods described earlier were then implemented to impute the missing values.

4.2 Performance Measure

To evaluate the performance of the aforementioned methods in imputing missing values, we utilised the Normalised Root Mean Square Error (NRMSE). The NRMSE is given by

$$NRMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n\sigma^2}}$$

where y_i represent the observed values in the data, $\hat{y}_i$ represents the imputed values, σ^2 is the variance of the observed data, and n is the number of observed data in each variable.

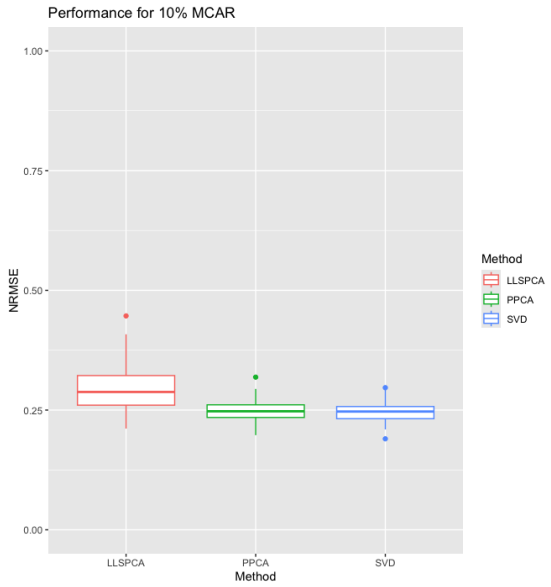
NRMSE values closer to 0 will indicate higher accuracy of the imputation

method, while higher values indicate lower accuracy of the imputation method.

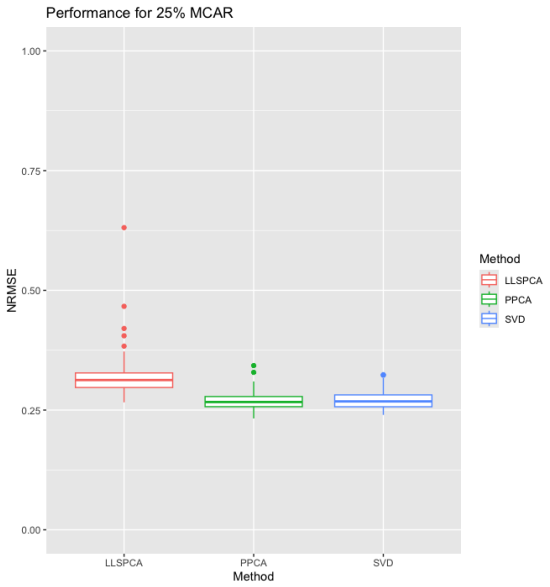
4.3 Results for MCAR and MAR for different proportions of missingness

Different percentages of missing values (10 %, 25 %, 40 %, 55%) were introduced into the data using `missMethods` package under the assumption of Missing Completely at Random (MCAR) and Missing at Random (MAR). Each of the three methods were implemented in R for 100 iterations and the Normalized Root Mean Squared Errors (NRMSE) were calculated and stored. For SVDPCA, we chose $k = 3$ as the most significant PCs, which accounted for 86% of the variation in the original data. For LLSPCA, different k values were implemented to determine the optimal k that will lead to higher performance of the method. We found $k = 3$ as the optimal number of nearest neighbors. In the figures below, the results are visualized. The mean and standard deviations of the NRMSE of each method are presented in Tables 1 and 2 in the appendix. Figure 1 displays the performance of the 3 methods for different percentages of MCAR values. In this figure, the median NRMSE is highest for the LLSPCA, followed by PPCA and then SVDPCA. This pattern is consistent for 10% and 25% missingness, though the NRMSE values are larger across all methods for 25% than for 10%. However, for higher proportions of missing values (40% and 55%), though the LLSPCA still has the largest median NRMSE, the PPCA outperforms the SVDPCA.

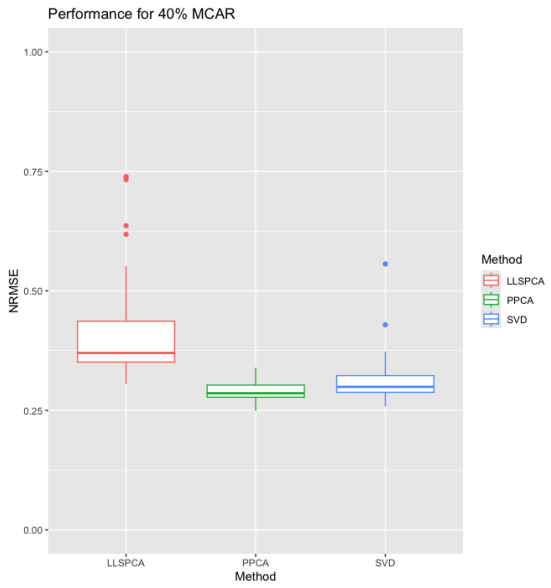
These patterns are also noticed in the NRMSE values for the Missing at Random (MAR) scenario in Figure 2 , where LLSPCA has the largest median NRMSE, but the PPCA outperforms SVDPCA from 40% missing values and beyond.



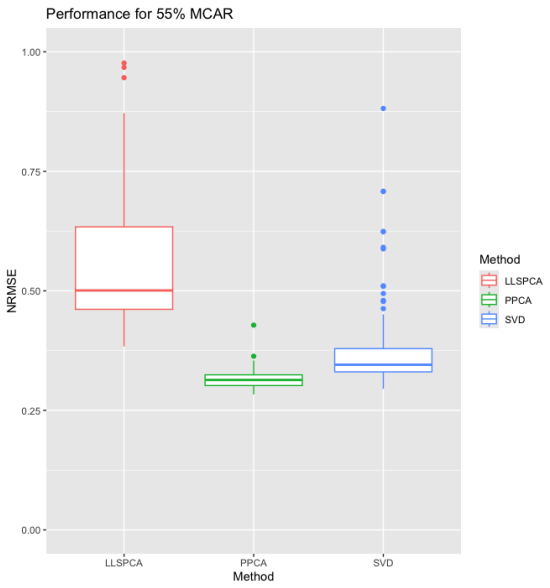
(a) 10% MCAR values



(b) 25% MCAR values

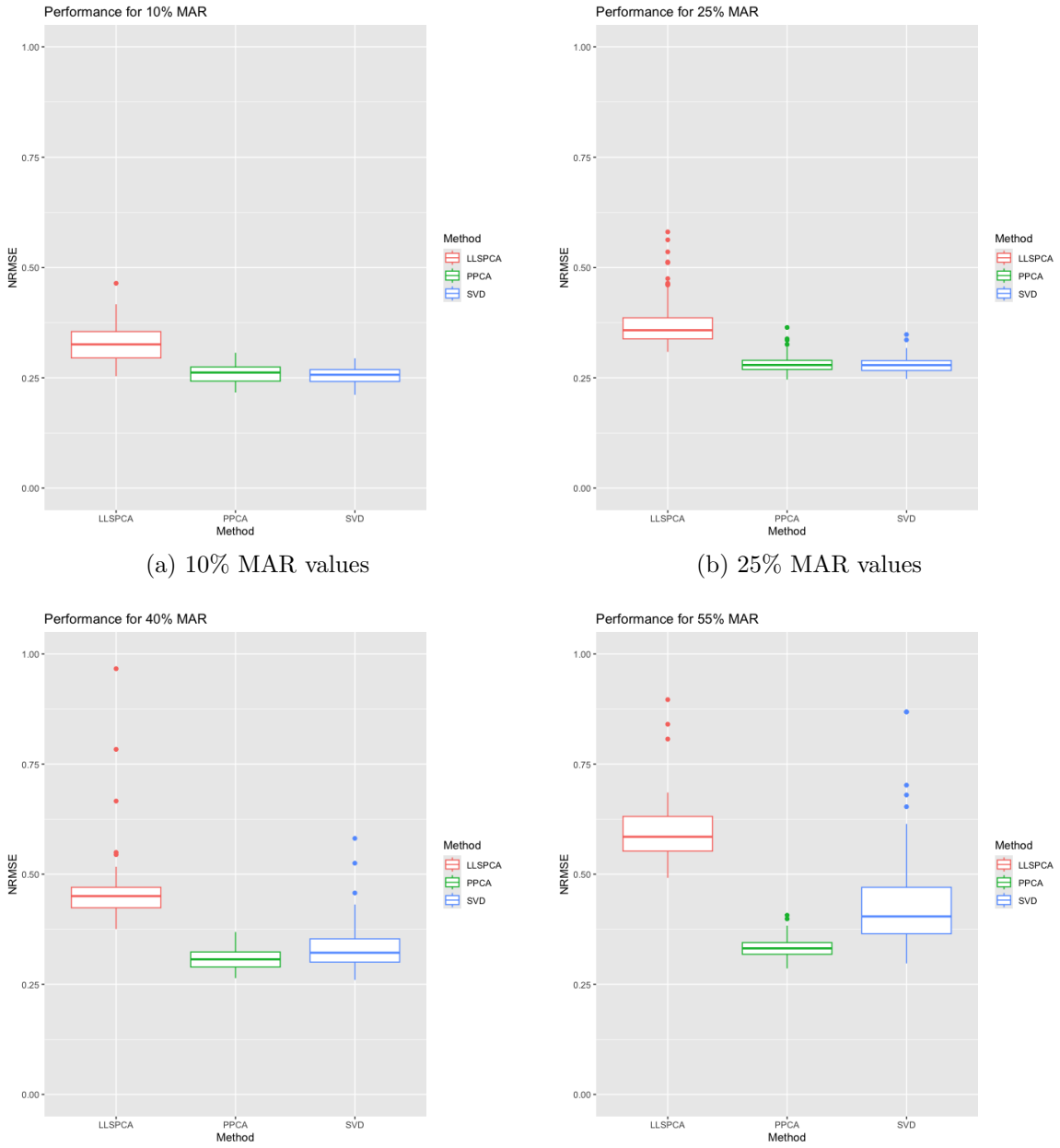


(c) 40% MCAR values



(d) 55% MCAR values

Figure 1: Boxplots showing the Normalized Root Mean Square Error of the methods for MCAR values at different levels of missing values.



(a) 10% MAR values

(b) 25% MAR values

(c) 40% MAR values

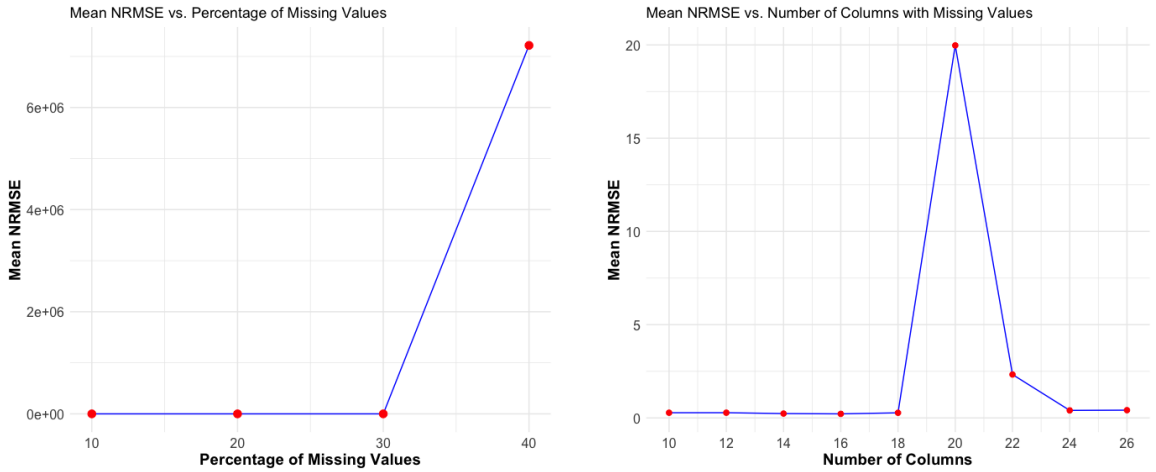
(d) 55% MAR values

Figure 2: Boxplots showing the Normalized Root Mean Square Error of the methods for MAR values at different levels of missing values.

It is worth noting that the LLSPCA sometimes led to excessively high NRMSE values for all percentages of missing data. Due to the significant impact of these values, they were excluded from the dataset for the purpose of visualisation. We excluded NRMSE values greater than 1.5. For MCAR, 4 values were excluded for 10%, 15 for 25%, 58 for 40% and 55 for 55%. For MAR, 3 values were excluded for 10%, 7 for 25%, 17 for 40% and 67 for 55%. In the next section, we investigate possible factors contributing to the poor performance of this method.

4.4 Local Least Squares PCA Challenges

From the results, it was shown that the LLSPCA method performed poorly, compared to the other methods. To investigate the factors contributing to the performance of LLSPCA, we investigated how the percentage of missingness as well as the the number of columns which contain missing values influence the performance of this method. We explored these for the Missing at Random (MAR) case and each method was implemented for 100 iterations. The results from these investigations are shown in the plots below.



(a) LLSPCA up to 40%

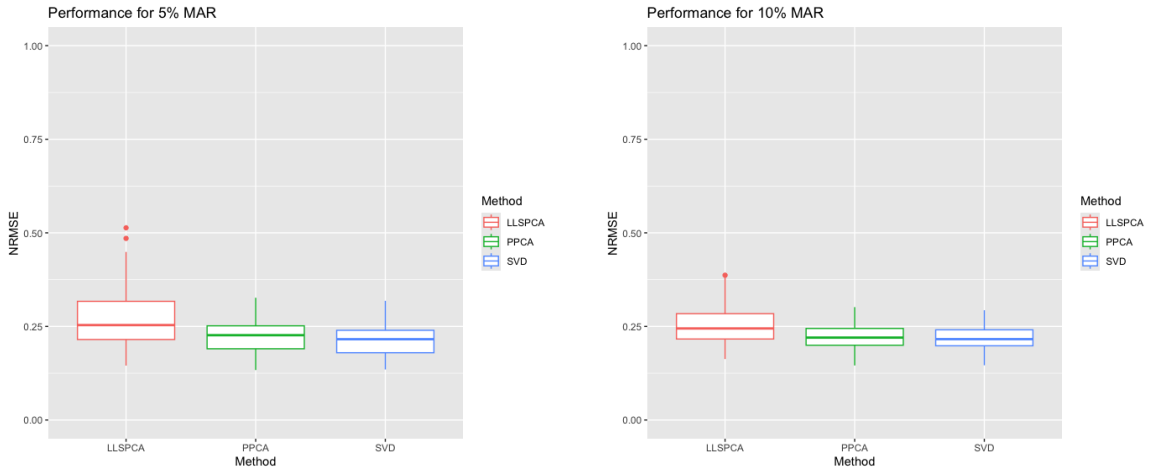
(b) LLSPCA for 20% MAR in different columns

Figure 3: Plots of Mean NRMSE for LLSPCA across different percentages of missing values and missing values in specified number of columns of the data over 100 iterations.

From figure 3(a), we discovered that LLSPCA performs poorly for missing values beyond 30%. The mean NRMSE values for each percentage are displayed in table 5 in the Appendix. Using this information, missing values were introduced into a specified number of columns and this was done for 20% missingness. The results are displayed in figure 3(b), for a specific seed value in R. However, in general, it was found that that the NRMSE values become unstable when there are missing values in 17-18 columns or more. Based on this, the performance of LLSPCA was evaluated with the other methods for missing values that are at or below 20% and only in 16 columns. The results will be presented in the next section.

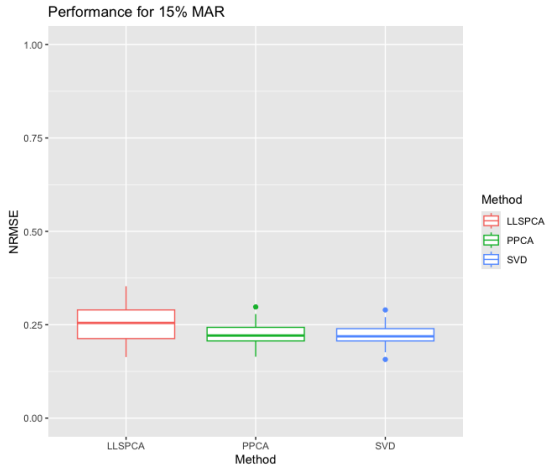
4.5 Updated Results

Figure 4 now shows that the LLSPCA, although still with a larger median NRMSE than the two methods, is significantly lower. It is also with noting that no values were excluded for these visualizations, as the NRMSE values across all 100 iterations were within reasonable range. These findings imply that, for this data, the number of columns with missing values as well as the overall percentage of missing values can

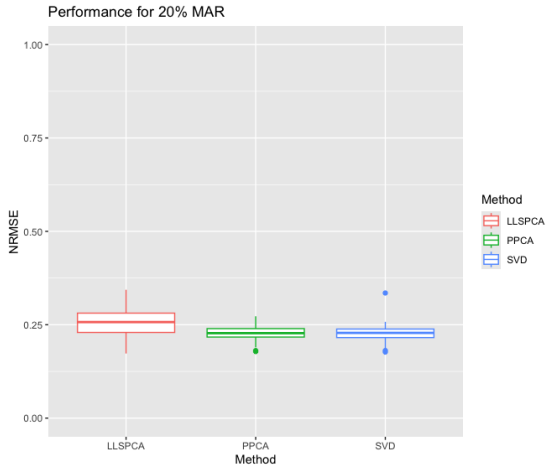


(a) 5% MAR values

(b) 10% MAR values



(c) 15% MAR values



(d) 20% MAR values

Figure 4: Boxplots showing the performance of the methods after introducing missing values to 16 columns of the data. Each method was run over 100 iterations.

impact the performance of the LLSPCA method. The SVDPCA and PPCA methods have results consistent with the results in figure 2, where the SVDPCA shows slightly lower median NRMSE for lower percentages of missingness (figure 4a and4b), but as the percentage of missingness increases, the NRMSE for both SVDPCA and PPCA become roughly the same (figure 4c and 4d).

5 Discussion

In this project we explored the performance of Singular Value Decomposition PCA, Local Least Squares PCA and Probabilistic PCA methods in imputing values for Missing Completely at Random (MCAR) and Missing at Random (MAR) data. Overall, we found that across all methods, the mean NRMSE values are higher for the MAR case than for the MCAR case, as expected. Additionally, SVDPCA and PPCA have similar results for lower percentages of missing values (specifically up to 20%), implying that either of these methods will be appropriate for imputation of missing values for lower percentages of missingness. These results hold for both MCAR and MAR cases. However, for higher percentages of missing values (beyond 20%), the PPCA method performs better than SVDPCA, implying that PPCA method is more appropriate for imputing missing values for higher percentages of missingness.

LLSPCA had the highest NRMSE across both MCAR and MAR cases. Our investigation into the reasons for unusually high NRMSE values revealed that the percentage of missing values in the data, as well as the number of columns with missing values can impact the performance of this method. For this dataset, we found that LLSPCA performs better when the percentage of missingness is less than 30%, and if the missing values are in 16 or less columns out of 27 total columns. After implementing all three methods with this update, we found that the LLSPCA performance improved drastically, even though the NRMSE was still higher than that of the other two methods.

LLSPCA is a method that is found to perform well in other papers, but based on this project, that may not be the case for all datasets. If this method is implemented on a dataset and the outcomes are similar to the findings of this project, it can be worthwhile to further investigate how the percentage of missing values and the number of variables (columns) with missing values impact its performance.

References

Everitt, B. and Hothorn, T. (2011). *An introduction to applied multivariate analysis with R*. Springer Science & Business Media.

Hyun, K. (2013). The prevention and handling of the missing data. *Korean J Anesthesiol*, 64(5):402–406.

John, C., Ekpenyong, E. J., and Nworu, C. C. (2019). Imputation of missing values in economic and financial time series data using five principal component analysis (pca) approaches. *Central Bank of Nigeria Journal of Applied Statistics*.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer.

Kim, H., Golub, G. H., and Park, H. (2004). Missing value estimation for dna microarray gene expression data: local least squares imputation. *Bioinformatics*, 21(2):187–198.

Li, P., Stuart, E. A., and Allison, D. B. (2015). Multiple imputation: A flexible tool for handling missing data. *JAMA*, 314(18):1966.

Nguyen, T., Ly, H. T., Riegler, M. A., Halvorsen, P., and Hammer, H. L. (2023). Principal components analysis based frameworks for efficient missing data imputation algorithms. In *Asian Conference on Intelligent Information and Database Systems*, pages 254–266. Springer.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 1*, 2:559–572.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rockel, T. (2022). *missMethods: Methods for Missing Data*. R package version 0.4.0.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489.

Stacklies, W., Redestig, H., Scholz, M., Walther, D., and Selbig, J. (2007). pcamethods – a bioconductor package providing pca methods for incomplete data. *Bioinformatics*, 23:1164–1167.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):611–622.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525.

Yoon, D., Lee, E.-K., and Park, T. (2007). Robust imputation method for missing values in microarray data. *BMC Bioinformatics*, 8(S2).

Appendix

Method	**10%**	**25 %**	**40 %**	**55 %**
LLSPCA	0.6687257	5.469468	4.864967×10^9	286256.2
	(2.725017)	(35.12486)	$(4.8649678 \times 10^{10})$	(2555651)
PPCA	0.2365183	0.2462316	**0.2510717**	**0.2606672**
	(0.03008588)	(0.02150488)	(0.02066219)	**(0.01690271)**
SVDPCA	**0.232038**	**0.2438552**	0.2514775	0.2607766
	(0.02793949)	**(0.02032547)**	**(0.01922548)**	(0.01731296)

Table 1: Mean Normalized Root Mean Square Error of the 3 methods and (standard deviations) for different levels of values Missing Completely at Random (MCAR). The lowest means and standard deviations are in bold.

Method	**10%**	**25 %**	**40 %**	**55 %**
LLSPCA	6.113699×10^{13}	2.955020×10^{11}	4.070795×10^{11}	2.074265×10^{11}
	$(6.113699 \times 10^{14})$	(2.955018×10^{12})	(2.241×10^{12})	(1.51698×10^{14})
PPCA	0.259924	0.2812591	**0.24115**	**0.3338465**
	(0.02032932)	(0.02010198)	**(0.02228307)**	**(0.02210731)**
SVDPCA	**0.255321**	**0.2801688**	0.3330018	0.4349711
	(0.01907295)	**(0.01836679)**	(0.04810779)	(0.1050797)

Table 2: Mean Normalized Root Mean Square Error of the 3 methods and (standard deviations) for different levels of values Missing at Random (MAR). The lowest means and standard deviations are in bold.

Method	**5%**	**10 %**	**15 %**	**20 %**
LLSPCA	0.2722258	0.25022	0.2514673	0.2559868
	(0.08160983)	(0.04912931)	(0.04465657)	(0.03725447)
PPCA	0.2227678	0.2247663	0.2243388	0.2271396
	(0.04320312)	(0.03277379)	(0.02487691)	**(0.01954769)**
SVDPCA	**0.2162115**	**0.2197251**	**0.222001**	**0.2268008**
	(0.04277343)	**(0.0309615)**	**(0.02405244)**	(0.0215927)

Table 3: Updated Average Normalized Root Mean Square Error and (standard deviations) of the 3 methods for different percentages of values Missing at Random (MAR), after introducing missingness to only 16 columns and percentages less than 20%. The lowest means and standard deviations are in bold.