John, Chisimkwuo; Ekpenyong, Emmanuel J.; Nworu, Charles C.

**Article**

# Imputation of missing values in economic and financial time series data using five principal component analysis approaches

CBN Journal of Applied Statistics

**Provided in Cooperation with:**
The Central Bank of Nigeria, Abuja

*Suggested Citation:* John, Chisimkwuo; Ekpenyong, Emmanuel J.; Nworu, Charles C. (2019) : Imputation of missing values in economic and financial time series data using five principal component analysis approaches, CBN Journal of Applied Statistics, ISSN 2476-8472, The Central Bank of Nigeria, Abuja, Vol. 10, Iss. 1, pp. 51-73, https://doi.org/10.33429/Cjas.10119.3/6

This Version is available at:
https://hdl.handle.net/10419/219299

# Imputation of Missing Values in Economic and Financial Time Series Data Using Five Principal Component Analysis Approaches

**Chisimkwuo John[1], Emmanuel J. Ekpenyong[2] and Charles C.Nworu[3]**

*This study assesses five approaches for imputing missing values. The evaluated methods include Singular Value Decomposition Imputation (svdPCA), Bayesian imputation (bPCA), Probabilistic imputation (pPCA), Non-Linear Iterative Partial Least squares imputation (nipalsPCA) and Local Least Square imputation (llsPCA). A 5%, 10%, 15% and 20% missing data were created under a missing completely at random (MCAR) assumption using five (5) variables: Net Foreign Assets (NFA), Credit to Core Private Sector (CCP), Reserve Money (RM), Narrow Money (M1), Private Sector Demand Deposits (PSDD), from 1981 to 2019 using R-software. The five imputation methods were used to estimate the artificially generated missing values. The performances of the PCA imputation approaches were evaluated based on the Mean Forecast Error (MFE), Root Mean Squared Error (RMSE) and Normalized Root Mean Squared Error (NRMSE) criteria. The result suggests that the bPCA, llsPCA and pPCA methods performed better than other imputation methods with the bPCA being the more appropriate method and llsPCA, the best method as it appears to be more stable than others in terms of the proportion of missingness.*

**Keywords:** Financial time series; Imputation; Missing data; PCA

**JEL Classification:** C15, C32, C52, E29

DOI: 10.33429/Cjas.10119.3/6

## 1.0 Introduction

Working with financial time series does not require the data to have missing observations over a long period of time. This is because the statistical properties of the series are preserved by its sequence using such a complete data. Financial

[1]Department of Statistics, Michael Okpara University of Agriculture, Umudike, Nigeria. Email: chisi.john@yahoo.com

[2]Department of Statistics, Michael Okpara University of Agriculture, Umudike, Nigeria. Email: ekpesstat@yahoo.com

[3]Corresponding Author: Department of Statistics, Michael Okpara University of Agriculture, Umudike, Nigeria. Email: nccharles19@gmail.com

time series such as the financial stock market data, for various reasons, frequently contain missing values. The reason may be attributed to markets being closed for holidays, inability to capture financial data in the specified period of time, recording errors etc. Such missing data make it difficult to predict future stock prices using the most up-to-date market information (Sohae, 2015). Thus, if there is a disturbance in the sequence of the series in terms of observations, the problem of missing data arises, hence there is an urgent need to handle such problem. In time series, each record is unique; dropping it would leave us with a series with holes, unusable for many purposes (Tusell, 2005).

This is in contrast to one of the assumptions of the Box-Jenkins method which entails that the series be equally spaced over time and that there are no lost values in the series (Yaffee & McGee, 1999). If missing data exists either in a univariate or multivariate series, then carrying out an analysis with the series may not be possible (Yaffee & McGee, 1999). In addition, the time series plot will have a lot of holes and will look truncated. Therefore, in order to carry out an objective and 'neat' time series analysis, there is need to estimate (impute) and plug back those missing observations.

Vital and valuable information will be lost by discarding such observations or removing the corresponding cases when the number of missing values in the dataset is large. This may lead to selection bias. Similarly, the correlation structure of a dataset may not be captured if the decision is to plug missing values with zeros or with mean value over the samples. This method is far from optimal and the series properties may seriously be affected (Kerkri *et al.*, 2015).

A more advanced statistical models have been developed that can effectively impute missing data using information in the non-missing part of the dataset. But this is subject to the type of data, percentage of missing data, the missing data mechanism, correlation structure of the data, the distribution of missing entries in the data and the size of the data. One of such procedures is the principal component analysis (PCA) approaches (Armina *et al.*., 2017; Gautman and Ravi, 2015).

The works of Armina *et al.* (2017) have further stated how type of data affects the choice and performance of imputation models. They added that global methods such as svdPCA, bPCA, etc perform better on data sets with low entropy (data sets with low variability and less information and error), while local methods such as llsPCA, K-Nearest Neighbour (KNN), etc, perform better with high entropy data sets (sets with high variance with more information and error or noise). Moreover, data that obey some of the assumptions of imputation methods may cause such methods to be adopted in order to reduce bias and improve performance of such methods. In addition, methods that perform better when categorical data are used may not perform equally when continuous or interval data are used (see Schmitt *et al.*, 2015).

In the aspect of the effect of percentage of missingness on the kind of imputation method to be adopted, several authors have made significant contributions to this case. Little and Rubin (2002) emphasized that if the proportion of values missing is small, then such missing values should be ignored in analysis, as it would not have a significant effect on the results of the analysis. They further suggested 20% or less percentage of missingness as acceptable, although there does not appear to be a clear definition of how much data can be imputed. Moreover, Eekhout *et al.*(2014), in their study, showed that when a large percentage of subjects had missing values greater than 25%, multiple imputation methods performed better in multi-item variables.

When dealing with missing data, Siddique *et al.*(2012) stated that special concern must be given to the process that gave rise to the missing data, referred to as missing data mechanism. Most methods for generating Multiple Imputations assume the missing data mechanism is ignorable, where the probability that a value is missing does not depend on unobserved information such as the value itself (Siddique *et al.*, 2012). They further stated that, when data are non-ignorably missing, the probability that a value is missing does depend on unobserved information. Closely related to the concept of ignorability are the missing data mechanism taxonomies -: missing at random (MAR)', 'missing completely at random (MCAR)', and 'not missing at random (NMAR).' MAR requires that the probability of miss-

ingness depends on observed values only. MCAR requires that the probability of missingness does not depend on either the observed values or the missing data and MNAR requires that the probability of missingness could depend on the value of the variable. Schmitt *et al.* (2015) and Dray and Josse (2015) have shown that the performances of the multiple imputation methods are affected by the missing data mechanism. They also stated that advanced multiple imputation methods are adopted under MAR and MCAR mechanism.

The correlation structure of the data is also a major factor in determining the performance of some multiple imputation methods. The works of Dray and Josse (2015) clearly indicated nipalsPCA and mean methods of imputation poorly performed when all variables considered were highly correlated, while the Iterative PCA (IPCA) performs better when correlation structure between variables are stronger.

For the effect of sample size on the performances of multiple imputation methods, the study of Schmitt *et al.* (2015) indicated that among the various methods compared, Fuzzy K-Means (FKM) became more robust with a more significant advantage than bPCA when applied to small data sets, but for large data sets they perform almost equally.

Schmitts *et al.* (2015) confirmed that missing data introduce an element of ambiguity into data analysis. They affect properties of statistical estimations such as means, variances, percentages and parameters, resulting in a loss of power and misleading predictions, inferences and conclusions. This is why missing values in financial time series data could result in misleading inferences and conclusions after analysis. It is also of note that these variables differ in their distributions from country to country and earlier stated distributions of variables affect the method to be used in imputing any form of missing values. In view of the importance of these financial indicators or variables in the financial and monetary sector development in Nigeria, there is a dire need to compare some PCA imputation methods in order to ascertain the appropriate methods that are suitable with the Nigerian financial and monetary variables considered in this work, since they have distribu-

tions different from other countries'.

The objective of this work is to compare some PCA imputation procedures namely; Singular Value Decomposition (svdPCA) imputation, Probabilistic PCA (pPCA) imputation, Bayesian PCA (bPCA) imputation, Non-linear Iterative Partial Least Squares PCA (nipalsPCA) imputation and Local Least Squares PCA (llsPCA) imputation with a view to determine the best performed PCA imputation method unlike the traditional methods (listwise deletion method, mean method etc) proposed by other authors which does not take into cognizance the whole data in the matrix simultaneously.

The remaining sections of this study are designed as follows: section two shows the review of related works; section three describes the method of data analysis; section four presents the analysis of results, and lastly section five is the conclusion which discusses the policy implications of this paper.

## 2.0 Literature Review

### 2.1 Theoretical Framework

Principal Component Analysis (PCA) is defined by Everitt and Dunn (1999) as a method that reduces the dimensionality of a set of multivariate data. This is done by partitioning a set of uncorrelated variables which is a linear combination of the original dataset. The derivation of the new variables are arranged in descending order of importance where the first principal component accounts for a greater variation in the original data while the remaining variation is accounted by the second component.

Let the random vector $X^T = [X_1, X_2, ..., X_p]$ have the covariance matrix $\sum$ with eigen values $\lambda_1 \geqslant \lambda_2 \geqslant ... \geqslant \lambda_p \geqslant 0$

Consider the equation

$$Y_i = a_i^T X = a_{i1}X_1 + a_{i2}X_2 + ... + a_{ii}X_i \quad (1)$$

Using the formula $\sum_{Y_p} = cov(Y_p)$, we obtain

$$\text{var}(Y_i) = a_i^T \sum a_i, i = 1, 2, ..., p \tag{2}$$

$$\text{cov}(Y_i, Y_k) = a_i^T \sum a_k, i, k = 1, 2, ..., p \tag{3}$$

Wichern and Johnson (2007) states that principal components are uncorrelated linear combinations $Y_1$, $Y_2$,..., $Y_p$ whose variances are as large as possible. Let $\sum$ be the covariance matrix associated with the random vector $X^T = [X_1, X_2, ..., X_p]$

Let $\sum$ have the eigenvalue-eigenvector pairs $(\lambda_1, e_1)$, $(\lambda_2, e_2)$, ..., $(\lambda_p, e_p)$ where $\lambda_1 \geqslant \lambda_2 \geqslant ... \geqslant \lambda_p \geqslant 0$. Then the $i^{th}$ principal component is given by

$$Y_i = e_i^T X = e_{i1}X_1 + e_{i2}X_2 + ... + e_{ip}X_p, i = 1, 2, ..., p \tag{4}$$

with these choices

$$\text{var}(Y_i) = e_i^T \sum e_i = \lambda_i, i = 1, 2, ..., p \tag{5}$$

$$\text{cov}(Y_i, Y_k) = e_i^T \sum e_k = 0, i \neq k \tag{6}$$

If some $\lambda_i$ are equal, the choices of the corresponding coefficient vectors, $e_i$, and hence $Y_i$ are not unique.

Ilin and Raiko (2010) stated that PCA can be derived from a number of starting points and optimization criteria. The most important of these are minimization of the mean square error in data compression, finding mutually orthogonal directions in the data having maximal variances and de-correlations of the data using orthogonal transformations. They further added that in the data compression formulation, PCA finds a smaller dimensional linear representation of data vectors such that the original data could be reconstructed from the compressed representation with the minimum square error.

Ke*et al.* (2018) noted that PCA-based approaches have at least three merits in the domain of missing data imputation. First, it does not require strict assumptions such as the daily similarity, no continuous incompleteness of data points, and a large database. Second, the principal components remove the relatively trivial

details and make sure that only the major information is used for constructing the probabilistic distribution of the latent variables. Third, it simultaneously achieves the high imputing accuracy, acceptable speed, and robustness to abnormal data points in a broad range of missing data imputation issues.

Ke *et al.* (2018) also added that PCA-based missing data imputation methods formulate the relationship between original variables and latent variables in a PCA-based form, and then solve the problem with EM iterations. In this method, the probability distribution of the compressed information based on the original observed data is first estimated, and then reconstructing the missing data by the compressed information, which can also be viewed as latent variables.

Suppose that we have $m$ samples of $d \times 1$ original vectors $y_1, y_2, y_3, ..., y_m$, which can be formulated as a function of $c \times 1$ dimensional latent variables:

$$y_j = \mathbf{W} z_j + \mu \tag{7}$$

where $\mathbf{W}$ is a $d \times c$ matrix, $z_j$ is a $c \times 1$ vector of principal components (i.e latent variables) and $\mu$ is a $d \times 1$ bias term; for the Local Least Squares PCA (llsPCA), a straightforward method to determine the latent variables is to minimize the mean square error between the reconstructed $y_{ij}$ attained from latent variables and the original observed $y_{ij}$ :

$$\min \sum_{i,j \in o} (y_{ij} - \hat{y}_{ij})^2 \tag{8}$$

$$\hat{y}_{ij} = \mathbf{W}_i^T z_j + \mu_i = \sum_{k=1}^{c} w_{ik} z_{kj} + \mu_i \tag{9}$$

where $y_{ij}$ means the $i^{th}$ variable of the $j^{th}$ sample of the observed data, while $\hat{y}_{ij}$ is the reconstruction of the data element $y_{ij}$ .$O$ is the set of indexes $i$, $j$. $z_{kj}$ means the $k^{th}$ latent variable of the $j^{th}$ sample of the latent space.

However, Friedland *et al.* (2008) stressed that the optimization problem in (8) can be solved by a least squares algorithm which updates parameters $\mathbf{W}$, $\mu$ and $z_j$. Ke *et al.* (2018) further said that the llsPCA method might easily suffer from the

over fitting issue, especially when the missing ratio is high, since the objective is to minimize the mean square error; thus the method may generate unreasonable large parameters to fit well observed data and lose the generalization ability.

Shi *et al.* (2013) and *Ke et al.* (2018) explained the method of Probabilistic PCA (pPCA) as a natural solution to the over-fitting problem of the Local Least Squares PCA (llsPCA) by adding a regularization term in the objective function to penalize unreasonably large parameters. Another solution is altering the transformation between the original data and latent variables to a probabilistic form, from which the regularization term is naturally derived. pPCA is derived by adding an error or isotropic term to equation (7):

$$y_j = \mathbf{W}z_j + \mu + \varepsilon_j \tag{10}$$

where $z_j,\ \varepsilon_j$ follow the normal distributions, i.e., $z_j \sim N(0,1),\ \varepsilon_j \sim N(0, vI)$. There are three groups of parameters, i.e., $\mathbf{W}$, $\mu$ and $v$ which can be estimated by the EM algorithm (Bishop, 1999).

Probability PCA (pPCA) is sometimes sensitive to the initialization parameters $\mathbf{W}$, $\mu$ and $v$. To overcome this defect, an assumption of the Gaussian prior probabilistic distribution was made to parameters and which formulates the Bayesian PCA (bPCA).$\mathbf{W}$ and $\mu$ follow a normal distributions: $\mu \sim N(0, v_\mu I)$, $w_i \sim N(0, v_{w,k} I)$ where $v_\mu,\ v_{w,k}$ are hyperparameters that can be updated, (*Ke et al.*, 2018).

Yoon, *et al.*, (2007) showed that the methods of imputation earlier described have the limitation of taking care of data with multi-collinearity and outliers especially when they exist with small sample size. To overcome this problem, they made use of the Robust PCA (rPCA) method. Robust PCA (rPCA) makes use of the principal components, instead of the original data, during least squares estimation of parameters.

Another imputation method, the Non-linear Partial Least Squares PCA (nipalsPCA) method bridged the gap between the use of standard linear PCA by meth-

ods earlier reviewed and the non-linear generalization of standard linear PCA. It further introduced the use of Partial Least Squares estimation method for estimating the parameters involved in the imputation model, (Ping,*et al.*, 2014).

### 2.2 Empirical Literature

Schmitt *et al.* (2015) compared six methods of imputing data namely: Mean, K-nearest neighbor (KNN), Fuzzy K-means (FKM), Singular Value Decomposition (SVD), Bayesian principal component analysis (bPCA) and Multiple imputation by chained equations (MICE) using four different reference data-sets split into two groups of various sizes: small dataset and large dataset, under a missing completely at random (MCAR) mechanism. Performance accuracy were measured based on Root Mean Square error (RMSE), Unsupervised Classification Error (UCE) and Supervised Classification Error (SCE). They concluded that the bPCA and FKM performed better than the other four imputation methods. They further stated that FKM outperformed bPCA when small datasets were considered. On the other hand, they emphasized the effect of the type of data on the performance of the imputation methods. They concluded that theirs were matrices of numerical values (biological data) and that they did not consider longitudinal and nominal data. In this paper, we seek to apply it to financial and economic data, which are known to be volatile.

Juha (2011) considered two methods of handling missing values: Robust PCA imputation algorithm and the nearest neighbor method. The performance of these methods varied for the simulated datasets and real world forest datasets. He concluded that the nearest neighbor method seems to be more useful for real data but this depends on the correlation structure of the data. He further added that Robust PCA performed better where the data have outliers and unknown distributions. He emphasized that the performance of other methods could be influenced by outliers and the distribution of data.

Pedreschi *et al.* (2008) worked on different methods of handling missing values. They considered three approaches namely; (1) Nonlinear Iterative Partial Least Squares (nipalsPCA) imputation (2) K-nearest neighbor and (3) the Bayesian prin-

cipal component analysis (bPCA). They applied these techniques to two sets of data. From the three tested methods of handling missing values, they concluded that the bPCA imputation approach proved to be the most consistent. They also added that for the parametric methods of imputing data to be efficient in performance, the normal assumptions of normality and homoscedasticity must be met.

Brock *et al.* (2008) evaluated eight imputation methods on nine different datasets of various types and sizes; which includes multiple exposures, time series and mixed type of data. The objective of their work was to assess the performance of the estimation methods under different conditions and to recommend appropriate use of these methods. These methods were compared in terms of percentages of missing data and the imputation accuracy was measured using the root mean squared error (RMSE). Their results showed that the bPCA outperformed other methods on data with strong correlation structure.

Yoon *et al.* (2007) applied the robust PCA (rPCA) methods to impute microarray data and compared it with bPCA, Local Least Squares imputation (llsimpute) and K- Nearest Neighbour (KNN). Using the normalized root mean squared error, they observed that rPCA outperformed other methods, but competed favourably with bPCA. They further added that bPCA performed based on the number of principal components and the type of data. Moreover, the bPCA is computationally expensive in terms of time because of the EM algorithm involved.

Ping *et al.* (2014) compared eight imputation methods based on accuracy and stability. The methods were svdPCA, pPCA, bPCA, Nonlinear PCA (NLPCA), nipalsPCA, least squares imputation (llsimpute), MICE and Multiple imputation methods. They applied them on 20 clinical features – age, gender, limour number, the size of the maximal tumor, liver cirrhosis, Barcelona Clinic Liver Cancer Staging Classification and 14 Serum Laboratory tests. From their analysis, they observed that bPCA may not be a suitable method to impute missing values for developing predictive model, as it could not achieve better performance than the complete data set.

Most of these works applied these imputation methods on medical and biological data, which are mostly count data and in some cases measurable; but we seek to apply different PCA methods on some financial or economic time series data of Nigeria (which change in terms of volatility and distributional structures from country to country) to evaluate their performances.

## 3.0 Methodology

### 3.1 Data Source

Data from quarterly monetary aggregates (N' billion) from first quarter of 1981 to first quarter of 2019 was obtained from Central Bank of Nigeria (CBN). Five variables which include Net Foreign Assets, Credit to Core Private Sector, Reserve Money, Narrow Money, and Private Sector Demand Deposits were used for the study. For the purpose of this work, values were made to be missing at random under an MCAR assumption at different percentages (5, 10, 15, 20) from the original dataset, using R-statistical package. Furthermore, the missing values were estimated using various PCA imputation techniques to recover the missing values referred as imputed data. Finally, the imputed data obtained was compared to their corresponding observed values.

### 3.2 PCA Imputation Approach and Model Specification

Five imputation methods used in this work are briefly discussed below. The methods assume that the data used satisfy the principal component analysis assumptions.

#### 3.2.1 Singular Value Decomposition (svdPCA) Imputation

The SVD imputation proposed by Troyanskaya *et al.* (2001) imputes missing data by a linear combination of a set of mutually orthogonal patterns to obtain an estimated value. In singular value decomposition, the $m \times n$ matrix, $m > n$ is expressed as a product of three matrices

$$Y = U \sum V^T \tag{11}$$

where $U$= $m \times m$ orthogonal matrix $V = n \times n$ orthogonal matrix $\sum = m \times n$

diagonal matrix.

The solution of the imputed missing value is improved iteratively until a total change in the matrix falls below a prescribed threshold, usually 0.01. This approach is known as the expectation-maximization (EM) approach. SVD impute seems to perform better on data with relatively high proportion of missing data, say 10% and above.

### 3.2.2 Probabilistic PCA (pPCA) Imputation.

As an extension to the traditional principal component analysis (PCA), pPCA is deduced with hidden variables by Gaussian model (Gaussian latent variable model). The generative model of pPCA as proposed by Tipping & Bishop (1999) is given by

$$t_i = W_{x_i} + \mu + \varepsilon_j \tag{12}$$

where $\varepsilon_j$ is d-dimensional vector of the noise,

$$\mu = \frac{\sum_{k=1}^{N} t_k}{N} \; isthesamplemean$$

$W$ is $d \times q$-dimensional parameter matrix and

$x \sim N(0, I_q)$, $\epsilon \sim N(n0, \sigma^2 I_d)$

PPCA works well on data with proportion of missing values between 10% and 15%. If the missing number of data exceeds this threshold, then the solution is likely not to converge.

### 3.2.3 Bayesian PCA (bPCA) Imputation

Similar to probabilistic principal component analysis (pPCA), the likelihood of an imputed value is obtained with the combination of the expectation maximization approach and the Bayesian estimation method (Stacklies *et al.*, 2007). The algorithm seems to be tolerant to relatively high proportion of missing data say, 10%. Oba *et al.* (2003) highlighted the three processes involved in estimating missing values using the bPCA. They are: Principal Component (PC) regression, Bayesian estimation and an expectation-maximization (EM)-like repetitive algorithm. The

posterior of the missing data based on the bPCA according to Oba et al. (2003) is given by:

$$q(Y^{miss}) = \int p\left(Y^{miss}\middle|\, Y^{obs}, \theta\right) q(\theta) d(\theta) \tag{13}$$

which corrreponds to the Bayesian PC regression.

### 3.2.4 Non-Linear Iterative Partial Least Squares (nipalsPCA) Imputation.

Another method of imputing missing data is the nipalsPCA imputation method. The nipalsPCA uses the elements of the principal component analysis of a finite dimensional random vector through a Jacobi-like iterative method (Tenenhaus, 1998). Here, NIPALS provides not only an estimation of principal factors and components, but also by the mean of the data reconstitution formula, an imputation method for missing data. The NIPALS has the advantage of working well in MAR cases (Cristian *et al.*, 2005). The NIPALS algorithm is easy to implement in standard programming languages.

Here, we introduce the NIPALS algorithm in the multivariate finite dimensional case. Let $Y = (Y_1, Y_2, \ldots, Y_p)^T$ be a random vector of dimension $p$, $p \geq 1$, such that $E(X_i) = 0$, $\forall i \epsilon 1, \ldots, p$. The expression of the vector $Y$ in terms of principal components and principal factors is a well-known result in multivariate data analysis (Escoufier, 1970). let

$$Y = \sum_{h=1}^{q} \xi_h U_h \tag{14}$$

where $q = dim\, L2(X)$, $\{\xi_h\}_{h=1,2,\ldots,q}$ and $\{U_h\}_{h=1,2,\ldots,q}$ are the principal components (random variables) and the principal factors of the principal analysis of $X$. It is worthy to note that, nipalsPCA is tolerant to small proportions of missing data, in most cases not more than 0.05. Also, for large data matrices or matrices that have a high degree of column collinearity, the orthogonality of the matrix is lost due to machine precision accumulated in each iteration step.

### 3.2.5 Local Least Squares PCA (llsPCA) Imputation

The llsPCA proposed by Kim et al. (2005) is based on the linear combination of the k-nearest neighbors of a missing dataset. The llsPCA is built on the Pearson correlation coefficient $r_{ij}$ (Pearson, 1894) which measures the strength of variables. The strength of the variables is measured as the absolute value of the distance between the variables. Where there is a missing value in the first position of a variable, say, $g_1$, $y_{ij}$ between two vectors $g_i' = (g_{i2}, \ldots, g_{in})^T$ and $g_j' = (g_{j1}, \ldots, g_{jn})^T$ becomes

$$y_{ij} = \frac{1}{n-1} \sum_{k=2}^{n} \left( \frac{g_{ik} - \overline{g}_i}{\sigma_1} \right) \left( \frac{g_{jk} - \overline{g}_j}{\sigma_j} \right) \tag{15}$$

where $\bar{g}_i$ and $\bar{g}_j$ are the mean values of $g_i'$ and $g_j'$ respectively and is the standard deviation of the values. In estimating missing values, the llsPCA performs better as the percentage of variables increases (Kim *et al.*; 2005).

## 3.3 Measures of Performance

The performance of PCA imputation methods were measured using the following methods:

1. Mean Forecast Error (MFE)
2. Root Mean Squared Error (RMSE)
3. Normalized Root Mean Squared Error (NMRSE)

**3.3.1 Mean Forecast Error (MFE):** The MFE according to Adhikari and Agrawal (2013) is given by

$$MFE = \frac{1}{n} \sum_{i=1}^{n} (y_{actual} - y_{imputed}) \tag{16}$$

$y_{actual}$ are the original values of the variables before they were made to miss, while the $y_{imputed}$ are the values imputed in places of the missing values.

It is a measure of the average deviation of forecasted values from actual ones. It shows the deviation of error and thus also termed as the Forecast bias. A zero MFE does not mean that forecasts are perfect, that is, contain no error, rather it only indicates that forecasts are on proper target. It depends on the scale of

measurement and also affected by data transformations.

**3.3.2 The Root Mean Squared Error (RMSE):** The RMSE (Adhikari and Agrawal, 2013) is given by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{actual} - y_{imputed}\right)^2} \tag{17}$$

It is a measure of average squared deviation of imputed values. RMSE -does not provide any idea about the direction of overall error. It gives an overall idea of the error occurred during forecasting. RMSE is a good measure of overall forecast but it is sensitive to change of scale and data transformation.

**3.3.3 The Normalized Root Mean Squared Error (NRMSE):** The NRMSE(Adhikari and Agrawal, 2013) is a balanced error measure and is very effective in judging accuracy of model. NRMSE is given by

$$NRMSE = \sqrt{\frac{mean[(y_{actual} - y_{imputed})^2]}{\text{variance}[y_{actual}]}} = \sqrt{\frac{1}{n\sigma^2_{actual}}\sum_{i=1}^{n}\left(y_{actual} - y_{imputed}\right)^2} \tag{18}$$

From the formula above, the variance of the actual value is calculated from the whole dataset and n is the number of samples in each variable in the whole data in the matrix. As the NRMSE value tends to 0.00, the more accurate the imputation method. But when the imputation method is too poor or when the noise associated with the data is too large, the NRMSE approaches to 1.00. NRMSE is not affected to change of scale and data transformations.

## 4.0 Analysis of Results

**Table 1:** Predictive accuracy of the five PCA imputation methods using the Mean Forecast Error at different levels of missingness

| Imputation Method | 5% | 10% | 15% | 20% |
|---|---|---|---|---|
| svdPCA | -2.16000 | 49.05000 | 52.82000 | -21.19000 |
| bPCA | -9.78000 | 21.22000 | 38.87000 | -40.95000 |
| pPCA | -10.18000 | 21.23000 | 39.86000 | -58.36000 |
| nipalsPCA | 1.58000 | 70.26000 | 37.72000 | 27.28000 |
| llsPCA | -5.45000 | 17.38000 | 27.28000 | -27.97000 |

Table 1 shows the performance indicators of the five PCA imputation methods based on the mean forecast error (MFE) at different percentage of missing data. The MFE indicators show the direction of the imputed values. The negative values indicate under-estimation while the positive values indicate over-estimation.

At 5% missingness, the nipalsPCA performed better than other imputation methods with its MFE value of 1.58 being closer to zero than other imputation methods; this was followed by the svdPCA, llsPCA, bPCA and pPCA with MFE values of -2.16, -5.45, -9.78 and -10.18 respectively. For 15% missingness, the llsPCA performed better than other imputation methods with its MFE values of 27.28, while in 20% missingness, the svdPCA performed better with MFE of -21.19. They were followed by nipalsPCA in both missingness with MFE of 37.72 and 27.28 respectively. However, at 10% missingness, the llsPCA performed better than other imputation methods with its MFE value of 17.38 being closer to zero than other imputation methods; this was followed by the bPCA, pPCA, svdPCA and nipalsPCA with MFE values of 21.22, 21.23, 49.05 and 70.26 respectively.

Generally, the llsPCA was observed to show consistency across the levels of missingness using the MFE except at 5% and 20% missing level where the svdPCA performed well.

**Table 2:** Predictive accuracy of the five PCA imputation methods using the Root Mean Square Error at different levels of missingness.

| Imputation Method | 5% | 10% | 15% | 20% |
|---|---|---|---|---|
| svdPCA | 199.97000 | 608.12000 | 603.47000 | 401.97000 |
| bPCA | 143.25000 | 370.64000 | 443.19000 | 362.92000 |
| pPCA | 149.99000 | 368.21000 | 444.65000 | 568.82000 |
| nipalsPCA | 555.44000 | 992.87000 | 960.97000 | 1191.57000 |
| llsPCA | 112.87000 | 396.86000 | 397.79000 | 358.14000 |

Table 2 shows the performance indicators of the five imputation methods based on the root mean square error (RMSE) at different percentage of missing data. The smaller the performance indicators, the more accurate the imputed values of the data.

**Table 3:** Predictive accuracy of the five PCA imputation methods using the Normalized Root Mean Square Error at different levels of missingness.

| Imputation Method | 5% | 10% | 15% | 20% |
|---|---|---|---|---|
| svdPCA | 0.04500 | 0.13600 | 0.13400 | 0.09000 |
| bPCA | 0.03200 | 0.08300 | 0.09900 | 0.08000 |
| pPCA | 0.03200 | 0.08300 | 0.09900 | 0.12700 |
| nipalsPCA | 0.12400 | 0.22100 | 0.21400 | 0.26500 |
| llsPCA | 0.02500 | 0.08800 | 0.08900 | 0.08000 |

For RMSE comparative measure, llsPCA performed better than other imputation methods in 15% of missingness considered with RMSE of 397.79, while pPCA took the lead for 10% missingness with RMSE of 368.21. For 5% and 20% missingness, llsPCA performed better with RMSE of 112.87 and 358.14 respectively. This was followed by bPCA. Therefore, using the RMSE as a means of performance measure, the llsPCA and bPCA also showed consistency across all the levels of missingness, but the pPCA performed better at 10% missingness.

The Normalized Root Mean Square Error (NRMSE) performance measures of the different imputation methods shown in Table 3, at different percentages of missing

data indicates that at 5% and 15% missingness, the llsPCA performed better than other imputation methods with their NRMSE values of 0.025 and 0.089 in each case. It was followed by bPCA and pPCA with the same NRMSE of 0.032 and 0.099 respectively. For 20% missingness, the llsPCA and bPCA performed equally better than other imputation methods with NRMSE value of 0.080. However, in 10% missingness, bPCA and pPCA performed equally better than others, which was followed by llsPCA. For Figure 1, the nipalsPCA shows deviation from zero indicating forecasts are out of point. The pPCA diverged from zero when the proportion of missing values increased. Also, the svdPCA, llsPCA and bPCA showed some level of consistency around zero indicating the imputed values are on point.

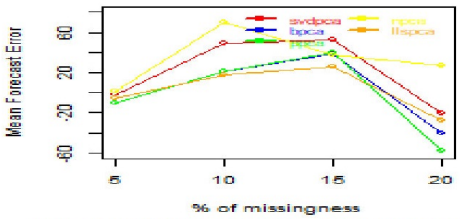


Figure 1: Predictive accuracy using MFE

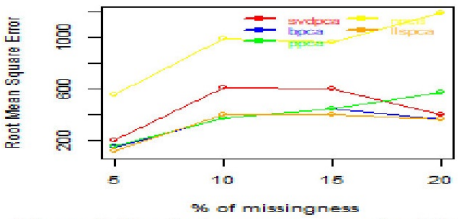


Figure 2: Predictive accuracy using RMSE

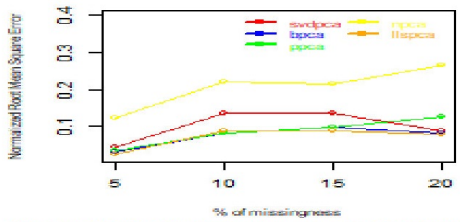


Figure 3: Predictive accuracy using NRMSE

Figures 1, 2 and 3 show the pattern of the performance of the different PCA methods based on the MFE, RMSE and NRMSE respectively. We can see that the three plots show the same behavior in pattern. The figures show that the performance of the methods largely depends on the percentage of missing values (5% to 20%) where a smaller value indicates a reliable estimation. Obviously, the performances deteriorated when the missing values increased and later decreased at 20% missingness for llsPCA, bPCA and pPCA. The llsPCA, bPCA and pPCA

outperformed other imputation methods. The nipalsPCA was the less effective imputation method where the difference was more pronounced when compared with other methods. The second worst method is the svdPCA.

## 5.0 Conclusion and Policy Implications

In this paper, we performed a comparison of five principal component analysis (PCA) imputation methods using the Nigeria quarterly monetary aggregates. First, we artificially created missing values from the original series under the assumption that it is missing completely at random (MCAR). The five methods include: singular value decomposition imputation (svdPCA), bayesian imputation (bPCA), probabilistic imputation (pPCA), non-linear iterative partial least squares imputation (nipalsPCA) and local least squares imputation (llsPCA). The choice of the best performed imputation method requires the consideration of three performance measures namely: mean forecast error (MFE), root mean square error (RMSE) and the normalized root mean square error (NRMSE). Most attention have been paid on the RMSE and NRMSE for measuring imputation accuracy but the use of the MFE has been undermined. Hence, we considered the MFE to determine the direction of the imputed value.

There was a deviation in the performance of these methods at 10% missingness using MFE measure in Table 1. Here, the svdPCA performed better than other methods and this may be due to the satisfaction of the normality assumptions of svdPCA. It may also be due to weakness of the MFE as a measure of performance, as such result did not indicate using othe performance measures. Also, in Table 2, For 5% and 20% missingness, llsPCA performed better with RMSE. This was followed by bPCA.. The performance of llsPCA may be a result of the variance structure of the data at 10% missingness as stated by Armina et al. (2017) that llsPCA performs well with data of high entropy data sets (sets with high variance with more information and error or noise) at 5% and 20% missingness. bPCA performed better than other imputation methods in 15% of missingness, while pPCA took the lead for 10% missingness using RMSE performance measure, but the difference is insignificant. This may be due to the fact that the normality and

homoscedasticity assumptions of pPCA were met in that data, (see Juha, 2011 and Yoon *et al.* 2007). From here, one could conclude that llsPCA, bPCA and pPCA may be applied to economic and financial data under the percentages of missingness. It is also worthy to note that in the literature, many authors have proposed different methods of imputation methods based on the correlation structure of data, type of data, distributions of data, type of missing mechanism, size of data and the proportion of missingness. Therefore, one can say that there is no general best method of imputation to be used.

In this paper, the three performance measures were consistent across the five imputation methods as shown in Figures 1, 2 and 3. The llsPCA, bPCA and the pPCA proved to be best performed method with the pPCA performance decreasing as the proportion of missing values increased. The significant performance of llsPCA may be due to a reason stated by Armina et al. (2017) that llsPCA performs well with data of high entropy data sets (sets with high variance with more information and error or noise). So, it is a well-known fact that financial data exhibit high level of volatility which brings about high variance and error. Also, the good performance of the bPCA was reported by Schmitt et al. (2015) in their comparison study where the imputation method confirmed better performance than other methods they used. But they did not compare their methods with other PCA imputation methods. The other methods they applied did not consider the correlation structure of the dataset. In conclusion, we recommend the llsPCA and bPCA imputation methods for estimating missing values in a financial time series data bearing in mind the proportion of missing values of the data. This may be due to the correlational structure of the financial series or the distributions exhibited by the data. Other methods, such as pPCA and svdPCA may also be recommended for imputing financial data in Nigeria, as they have good performance abilities based on their MFE, RMSE and NRMSE. The policy implication of the finding is that observations may be missing when predicting the Gross Domestic Product of Nigeria with respect to these variables; Net Foreign Assets, Credit to Core Private Sector, Reserve Money, Narrow Money, and Private Sector Demand Deposits using the time series or the regression approach. This study is useful for policy makers with information on the best method of imputation when it is clear

that there are missing observations in the dataset. It is obvious that the llsPCA and bPCA methods are preferred to the competing methods as observed in the results. These imputations would help policy makers to conveniently carry out economic analyses on the economic variables under consideration, with complete data sets. The predictions or inferences made on the basis of the analyses would further assist policy makers to formulate good and reliable policies, which could not have been done with missing or incomplete values of economic variables.

## References

Adhikari, R., & Agrawal, R.K (2013). An Introductory Study on Time Series Modeling and Forecasting. Retrieved from http://arxiv.org/abs/1302.6613.

Armina, R., Zain, A.M., Ali, N.A., & Sallehuddin, R. (2017). A review on missing value estimation using imputation algorithm. *Journal of Physics Conference Series*, 892(1): 012004.

Bishop, M.C. (1999). Bayesian PCA: Advances in Neural *Information Processing Systems. MIT Press*, 11, 382-388.

Brock, G.N; Shaffer, J.R; Brackesley, R.E; Lote, M.J; & Tsong, G.C. (2008). Which missing value imputation to use in expression profiles: A comparative study and two selection schemes. Journal of Bioinformantics, 9-12.

Cristian, P; Gilbert, S; & Mohammed, H.B.H.M. (2005).The NIPALS algorithm for missing functional data. *Romanian Review of Pure and Applied Mathematics*, 55(4), 1-12.

Dray, S & Josse, J. (2015). Principal component analysis with missing values: A comparative survey methods. *Journal of Plant Ecology*, 216, 657-667.

Escoufier, Y. (1970). Echantillonnagedansune population de variables aléatoiresréelles. *Rev. Roumaine Math Pures Appl*, 55, 315-326.

Everitt, B.S & Dunn, G. (2001). Applied Multivariate Data Analysis (2nd ed.). Chichester, WS: John Wiley & Sons, Ltd.

Friedland, S., Niknejad, A., Kaveh, M., & Zane, H. (2008). An algorithm for missing value estimation for DNA microarray data. In 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing - Proceedings, Vol. 2.

Gautrain, C., & Ravi, V. (2015). Data imputation via evolutionary computation, clustering and a neural network. Neurocomputing, 156:134-142.

Ilin, A., & Raiko, T. (2010). Practical approaches to principal component analysis in the presence of missing values. *The Journal of Machine Learning Research*, 11, 1957-2000.

Josse, J. (2015). Principal component analysis with missing values: A Comparative Survey of Methods. *Springer Science Business Media*, 216: 657-667.

Juha, K. (2011). Robust PCA methods for complete and missing data. Retrieved from http://users.ics.tkk.fi/juha.

Ke, J., Zhang, S., Yang, H., & Chan, X. M. (2018). PCA-based missing information imputation for real-time crash likelihood prediction under imbalanced data. Transportmetrica A: Transport Science, Volume 15, issue 2.

Kerkri, A, Zarrouk, Z. & Allal, J. (2015). A comparison of NIPALS algorithm with two other missing data treatment methods in a principal component analysis.

Kim, H; Golub, G.H; & Park, H. (2005). Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Journal of Bioinformantics*, 21(2), 187-198.

Oba, S; Sato, M.A; Takemasa, I; Monden, M; Matsubara, K; & Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. *Journal of Bioinformatics*, 19(16), 2088-2096.

Pedreschi, R; Hertog M.L; Carpentier S.C; Lammertyn, J; Robben, J; Noben, J.P; Panis, B; Swennen, R; & Nicolaï, B.M. (2008). Treatment of missing values for multivariatestatistical analysis of gel-based proteomics data. *Journal of Proteomics*, 8(7), 1371-1383.

Ping, X. O., Lai F., Tseng Y.J, Liang J. D., Huang G. T., Yang, P.M. (2014). Evaluation of imputation methods for missing data and their effect on the reliability of predictive models. International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies, 6, 8-14.

Schmitt, P; Mandel, J; & Guedj, M. (2015). A comparison of six methods for missing data imputation. *Journal of Biometrics & Biostatistics*, 6(224), 1-6.

Shi, F., Zhang, D., Chen, J., & Karimi, H.R. (2013). Missing values extraction for microarray data by Bayesian Principal Component Analysis and Iterative Local Least Squares. Mathematical Problems in Engineering, Vol. 2013, article ID 162938, pp. 1-5.

Siddique, J., Harel, O., & Crespi, C. M. (2012). Addressing missing data mechanism uncertainty using multiple-model multiple imputation: Application to a longitudinal clinical trial. *The Annals of Applied Statistics*, 6(4), 1814—1837.

Sohae, O. (2015). Multiple Imputation in missing values in time series data (Master's thesis). Duke University, North California.

Stacklies, W & Redestig, H. (2007). The pcamethods package. Retrieved from http://www.bioconductor.org/packages/level/bioc/vignettes/inst/doc/.

Tenenhaus, M. (1998). La régression PLS Théorieet pratique. In Rev. Roumaine Math Pures Appl, 55, 315-326. Retrieved from cedric.cnam.fr/fichiers/.

Tipping, M.E & Bishop, C.M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society*, 16(3), 611-622.

Troyanskaya, O; Cantor, M; Sherlock, G; Brown, P; Hastie, T; Tibshirani, R; Botstein, D; & Altman, R.B. (2001). Missing value estimation methods for DNA microarrays. *Journal of Bioinformatics*, 17(6), 520-525.

Tusell, F.P. (2005). Multiple imputation of time series with an application to the construction of historical price indices. Lehendakari Aguirre, Bilbao: University of the Basque Country Go.

Wichern, D.W & Johnson, R.A. (2007). Applied Multivariate Statistical Analysis (6th edition). Upper Saddle River, NJ: Pearson Prentice Hall.

Yaffee, R & McGee, M. (1999).Time Series Analysis and Forecasting: With Applications of SAS and SPSS. Brooklyn, NY: Academic Press Incorporation.

Yoon, D., Lee, E., Park, T. (2007). Robust imputation method for missing values in microarray data. BMC: *Bioinformatics*. DOI: 10.1186/1471-2105-8-S2-S6.