Journal of Theoretical and Applied Information Technology
15th October 2022. Vol.100. No 19
© 2022 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PRINCIPAL COMPONENT REGRESSION WITH VARIATIONAL BAYESIAN PRINCIPAL COMPONENT ANALYSIS APPROACH TO HANDLE MULTICOLLINEARITY AND MISSING DATA

NABILA AZARIN BALQIS[1], SUCI ASTUTIK[2], SOLIMUN[3], NURJANNAH[4], HENNY PRAMOEDYO[5]

[1]Student, University of Brawijaya, Department of Statistics, East Java, Indonesia

[2,3,4,5]Lecturer, University of Brawijaya, Department of Statistics, East Java, Indonesia

E-mail: [1]nabilaazarin24@gmail.com, [2]suci_sp@ub.ac.id, [3]solimun@ub.ac.id, [4]nj_anna@ub.ac.id, [5]hennyp@ub.ac.id

ABSTRACT

Principal Component Regression (PCR) is a combination method of Principal Component Analysis (PCA) and linear regression that aims to deal with multicollinearity in regression data. Classical PCA has the disadvantage that when faced with missing data. Missing data becomes a weakness in PCA where the resulting principal component will lose a lot of information so that the principal component cannot really describe the original variable properly. The method that can be used to deal with these problems and overcome overfitting is Variational Bayesian Principal Component Analysis (VBPCA). This study aims to modeling PCR using VBPCA with Ordinary Least Square (OLS) as a regression parameter estimation method to overcome multicollinearity at various levels of missing data proportions. The data used in this study are secondary data and simulation data which has been contaminated with collinearity in the predictor variables with various levels of the proportion missing data of 1%, 5%, and 10%. The results of this study indicate that in estimating the PCR parameters with VBPCA method using OLS, the estimated regression parameter coefficients have a constant value at the proportion of missing data up to 5%. This is influenced by missing data where the greater proportion of missing data, then the estimation results of the regression parameters are less constant and have a large standard error value of the regression parameters. Multicollinearity in secondary data and simulation data can be optimally overcome as indicated by the smaller standard error value of the regression parameter for the PCR method using VBPCA. VBPCA can handle the proportion of missing data to less than 10%. This is due to the large proportion of missing data as evidenced by the larger MAPE value, the cross validation (Q^2) and the adjusted R^2 value which are getting smaller as the proportion of missing data increases.

Keywords: *Missing Data, Multicollinearity, PCA, PCR, VBPCA*

1. INTRODUCTION

Multicollinearity is one of the problems that often occurs in multiple linear regression analysis caused by a strong correlation between predictor variables. The multicollinearity that occurs will cause the possibility that the least squares method of estimation cannot be obtained or can be obtained but will have a high variance of parameter estimators [1]. One method that can overcome multicollinearity in regression data is PCR. PCR is a combination analysis between linear regression analysis and PCA which is formed from two stages, namely forming the eigenvalues and eigenvectors of the sample covariance matrix which produces the principal component and then regresses to the response variable. PCR method has been proposed as an alternative to the OLS estimator when independent assumptions are not met in the analysis [2].

In field data, phenomena are often found where observations in the data have not been recorded or are missing, which is called missing data. Missing data is an observation on data that is not stored for a variable in the desired observation [3]. Problems that arise due to missing data is a significant influence on the conclusions that can be drawn from the when compared with complete data. Missing values in a data will cause the loss of information for each observation so that some research will focus on handling missing data first before carrying out further analysis. Missing data can result in bias, depending on the missing data mechanism and the statistical approach applied [4]. The problems caused by the missing data have an impact on the validity of the experiment and can lead to invalid conclusions [3].

Analytical methods in statistics are not entirely able to record and handle missing data. Some analyzes will return error results when there is missing data. In PCR, PCA can only handle multicollinearity on the predictor variables. However, it's ability to deal with missing data is incomprehensible and becomes an important open challenge [5]. Missing data is a weakness in PCA where the principal components produced will lose a lot of information so that the model cannot really describe the data properly. Missing data is common in medical research, but there is uncertainty about the acceptable limits of missing data when more complex methods are used, such as maximum likelihood, imputation, and Bayesian methods [6].

There are methods to handle missing data that are also contaminated with multicollinearity, namely Probabilistic Principal Component Analysis (PPCA) and VBPCA. VBPCA was introduced by Bishop in 1999 with the aim of selecting the number of components in PCA [7]. Then VBPCA was applied in PCA method with missing data [8]. The VBPCA method on missing data provides a high mean accuracy estimate [9]. Between the two methods, PPCA has a weakness that is prone to overfitting when the proportion of missing data is large. The overfitting will cause the principal components formed to be influenced only by certain observations so that additional information is needed through the prior density in the Bayesian Framework which assumes all parameters will be considered as random variables. Handling the overfitting problem can be handled by the VBPCA method in a Bayesian Framework and using a variational expectation-maximization iterative algorithm to search for each main sub-space, then automatically select the optimal number of principal components for data reconstruction [10].

Several studies applying VBPCA to missing data have been carried out. Yordani's research [11], which conducted an analysis of the principal components of missing data using the VBPCA method. The data used in his research is simulation data and it concluded that the VBCPA method can provide a high tolerance for missing data. In addition, Li's research [10], conducted a study on extracting common mode errors from regional Global Navigation Satellite System (GNSS) based on time series in the presence of missing data using VBPCA. The study concluded that the VBPCA method is more efficient alternative method for extracting CME general errors from the time series of regional GNSS positions in the presence of missing data. However, research on PCR using VBPCA as a principal component analysis method to determine the limits of missing data in overcoming multicollinearity and missing data has not been found so far.

According to Taufiq et al. [12] conducted a study that focused on Geographically Weighted Regression in Cox Survival Analysis for Weibull Distributed Data with Bayesian Approach. Meanwhile, according to Fernandes & Solimun [13] carried out research developments that implemented PCA in the innovation strategy by utilizing technology. Research by Raharjo et al [14] implements technological developments on organizational culture and job design using principal component analysis in the analysis of research variables indicators.

Based on the problems that have been described and based on previous research, this study was conducted with the aim of modeling PCR with the VBPCA approach using OLS to overcome multicollinearity and missing data. In addition, this study also aims to see the effect of the proportion of missing data on the regression model to determine the tolerable limit of missing data. The data used in this study are secondary data and simulation data generated from secondary data. The reason for choosing the simulation data is based on previous research regarding the proportion of missing data with a high degree of collinearity which is rarely found. Therefore, in this study, a simulation study was used to facilitate researchers in adjusting data that met the criteria of the model to be developed.

2. LITERATURE REVIEW

2.1 Missing Data

Missing data is defined as the value of the observation on the data that is not stored for a variable in the desired observation [15]. Statistical analysis tends to be biased when more than 10% of data are missing [16]. Missing data itself is common in medical research [17]. Most researchers assume that missing data does not interfere with data set analysis intrinsically [18], but it becomes even more critical when missing data concerns a multi-item instrument, as the lack of information even on one of its items leads to an inability to calculate total instruments scores [19]. Missing data can occur due to various factors such as completely random missing at random, missing not at random or accidental missing which may be the result of a system malfunction during data collection or human error during data pre-processing [20]. These factors are called the missing data mechanism. Mechanisms of missing data and patterns of missing data have a greater impact on research results than the proportion of missing data [21].

There are three types of missing data according to assumptions based on the missing data mechanism, namely Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR) [22]. In this study used MCAR. MCAR is a missing data mechanism that often occurs in field data. The MCAR method assumes that the data set is independent of unobserved and unobserved values [23]. The pattern occurs unsystematically and can be considered as a random subsample of the hypothesized data when complete. Missing data in the MCAR mechanism due to things such as equipment failure, samples lost in transit, and failures in the study design.

2.2 Missing Data Mechanism Assumption Test

Missing data mechanism assumption test is used to see whether the research data meets the missing data mechanism or not. In this study, the MCAR mechanism was used in the simulation data, so the assumption test for the MCAR mechanism used was Little's test of MCAR. The hypothesis for Little's test is given as follows [24]:

H_0 : $\boldsymbol{y}_{o,i}|\boldsymbol{r}_i \sim N\left(\boldsymbol{\mu}_{\boldsymbol{o}_j}, \Sigma_{\boldsymbol{o}_j}\right)$ (missing data pattern following the MCAR mechanism), *vs*

H_0 : $\boldsymbol{y}_{o,i}|\boldsymbol{r}_i \sim N\left(\boldsymbol{v}_{\boldsymbol{o}_j}, \Sigma_{\boldsymbol{o}_j}\right)$ (missing data pattern does not following the MCAR mechanism)

The test statistic for Little's MCAR is given in the following equation:

$$d^2 = \sum_{j=1}^{J} n_j \left(\overline{\boldsymbol{y}}_{\boldsymbol{o}_j} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{o}_j}\right)^T \Sigma_{\boldsymbol{o}_j}^{-1} \left(\overline{\boldsymbol{y}}_{\boldsymbol{o}_j} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{o}_j}\right) \quad (1)$$

where $\overline{\boldsymbol{y}}_{\boldsymbol{o}_j}$ with dimension $p_j \times 1$ is the average of the observed j-th missing pattern samples, $\boldsymbol{\mu}_{\boldsymbol{o}_j}$ and $\Sigma_{\boldsymbol{o}_j}$ are the average vectors with dimension $p_j \times 1$ and the covariance matrix with dimension $p_j \times p_j$ of the component observed only for the j-th missing pattern. The rejection area for Little's MCAR test statistic is the null hypothesis is rejected if $d^2 > \chi^2_{db}(1-\alpha)$, where α is the level of significance.

2.3 Non-Multicollinearity Assumption Test

Non multicollinearity assumption is an assumption which states that there is no correlation between two or more predictor variables in the data. Multicollinearity is a serious problem when they exist in econometric data, which arises as a result of violating the assumption of equal variance between the error terms and the independence between the predictor variables of the model. With this assumption, OLS violation will not provide the best linear model that is an unbiased, efficient and consistent estimator [25].

One way that can be used to detect the presence or absence of multicollinearity is to look at the Variance Inflation Factor (VIF) value of each predictor variable. The VIF value is determined to explore the level of multilinear relationship that exists between the predictor variables where each predictor variable is regressed against the remaining predictor variable as response variable [26]. The VIF value can be obtained by formula in the following equation:

$$VIF = \frac{1}{1-R_j^2} \quad (2)$$

where R_j^2 is the coefficient of determination of the auxiliary regression. Auxiliary regression is a regression with X_j as predictor variable and other X as response variables. The formula for R_j^2 is given in the following equation:

$$R_j^2 = \frac{SS_R}{SS_T} = \frac{\sum_{i=1}^{N}\sum_{i=1}^{T}(\hat{X}_{it}-\bar{X}_{it})^2}{\sum_{i=1}^{N}\sum_{i=1}^{T}(X_{it}-\bar{X}_{it})^2} \quad (3)$$

where N is multiple subjects, T is a number of observations of each subject, and p is a number of predictor variables.

The coefficient of determination can be determined by modelling one predictor variable with another predictor variable. In this study, referring to the VIF criteria, if there is a VIF > 10 then the correlation between the predictor variables is very high and the opposite also applies [27].

2.4 Linear Regression Analysis

Linear regression analysis is a method that is useful for determining the linear relationship of response variable with one or more predictor variables. One of the goals of regression analysis is to determine a good regression model that can be used to explain and predict things related to the variables involved in the regression model [28]. The linear regression model in general can be written as follows:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}) \tag{4}$$

with $\boldsymbol{Y}$ is a matrix of response variable, $\boldsymbol{X}$ is a matrix of predictor variables, $\boldsymbol{\beta}$ is a matrix of regression coefficient, and $\boldsymbol{\varepsilon}$ is a matrix of regression model error.

The regression parameter estimator that is often used is OLS. In matrix notation, the least squares method is the same as minimizing $\boldsymbol{\varepsilon'\varepsilon}$ with the equation [29]:

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X\beta} \tag{5}$$

therefore,

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta} = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta} \tag{6}$$

where $\boldsymbol{\beta'X'y}$ is a 1×1 matrix or a scalar. The reciprocal of $(\boldsymbol{\beta'X'y}) = \boldsymbol{y'X\beta}$ is a scalar. The least squares method estimator must satisfy:

$$\frac{\partial \sum_{i=1}^{n} \varepsilon_i^2}{\partial \boldsymbol{\beta}} = \frac{\partial \boldsymbol{\varepsilon\prime\varepsilon}}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{0} \tag{7}$$

when simplified the equation will become:

$$\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y} \tag{8}$$

to solve this equation, we multiply both by the inverse of $\boldsymbol{X'X}$. So, the OLS estimator of $\widehat{\boldsymbol{\beta}}$ as follows:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{9}$$

2.5 Classical Principal Component Regression

PCR analysis is a combination analysis between regression analysis and PCA. PCR analysis is a regression of the response variables to uncorrelated principal components, where each principal component is linear combination of all predictor variables that have been specified from the start [30]. The existence of correlation between predictor variables causes one of the basic assumptions of multiple linear regression analysis in OLS to fail to be fulfilled and one way to free the correlation between predictor variables is by PCR method. The new variable as the principal component (Q) is the result of the transformation of the predictor variable (X) whose model is in the form of a matrix in the linear regression equation and the principal component is written with the principal analysis equation where the weighting vector $\boldsymbol{a}_j^T$ is obtained by maximizing the variance of principal components, namely $\boldsymbol{S}_{\boldsymbol{Q}_j}^{\boldsymbol{2}} = \boldsymbol{a}_j^T \boldsymbol{S} \boldsymbol{a}_j$ with constraints $\boldsymbol{a}_j^T \boldsymbol{a}_j = 1$ and $\boldsymbol{a}_h^T \boldsymbol{a}_j = 0$, for $h \neq j$. The weighting vector $\boldsymbol{a}_j^T$ is obtained from the covariance matrix estimated with the matrix $\boldsymbol{S}$, namely $\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$. The vector $\boldsymbol{a}_j^T$ that satisfies the above constraints is the feature vector of the covariance matrix. The PCR model can be written as follows:

$$Y = \beta_0 + \beta_1 Q + \beta_2 Q_2 + \ldots + \beta_k Q_k + \varepsilon, \text{ with } j \leq k \tag{10}$$

2.6 Principal Component Regression with VBPCA Approach

PCA is a multivariate approach that is used to convert several correlated variables into linearly uncorrelated variables which can be called principal components [31]. The principal component itself is a linear combination of variables that have the largest number of variances and followed by the second, third, and so on [32]. PCA performed on the regression model is called PCR. In this study, PCR modelling was carried out using the PCA approach using the VBPCA method. The VBPCA method approach in PCR is intended to overcome multicollinearity in regression data that has missing observation vales. In the process, the principal components generated by VBPCA will be regressed using OLS.

VBPCA method uses the Expectation Maximization (EM) approach combined with the Bayesian estimation method to calculate the probability of the estimated value. VBPCA produces a weighting of factors to estimate the incomplete value to be different from the usual PCA technique, but the estimation of the incomplete value is getting better. The estimated missing value of the VBPCA method is given in the following equation [11]:

$$q(\boldsymbol{Y}^*) = \int d\boldsymbol{\theta}\, q(\boldsymbol{\theta}) p(\boldsymbol{Y}^* | \boldsymbol{Y}^{obs}, \boldsymbol{\theta}) \tag{11}$$

where $\boldsymbol{\theta}$ is a posterior parameter, $\boldsymbol{Y}^{obs}$ is a complete data set, and $\boldsymbol{Y}^*$ is an incomplete data set. The application of the VBPCA algorithm consist of several steps as follows [11]:

1. Perform principal component analysis by estimating missing values $\boldsymbol{Y}^*$.
2. Determines the conjugate prior distribution.
3. Initiates the posterior distribution of missing values $q(\boldsymbol{Y}^*)$ by imputing missing values.
4. Obtaining the posterior distribution of $\boldsymbol{\theta}$ parameters and $q(\boldsymbol{\theta})$ based on $\boldsymbol{Y}^{obs}$ observation data and based on the new posterior distribution of missing values $q(\boldsymbol{Y}^*)$.
5. The posterior distribution of missing values $q(\boldsymbol{Y}^*)$ then estimated based on the new posterior distribution of $q(\boldsymbol{\theta})$.
6. The hyperparameter $\boldsymbol{\alpha}$ then updated based on $q(\boldsymbol{\theta})$ and $q(\boldsymbol{\theta})$.
7. Repeat the process of steps 4 to 6 until it

converges.

VBPCA algorithm above will produce an estimate of all the missing values. So based on this process, the missing values has been resolved. After obtaining missing values estimates, complete observations are formed that can be used for further analysis. Estimators based on principal components that explain variations in the data. From the complete data that has been estimated, then steps are taken to find the principal components of the predictor variable group. The principal component is formed to overcome the multicollinearity of the predictor variables. In the process of finding the principal components, a covariance matrix is needed to form the eigenvalues and eigenvectors. The covariance matrix of the VBPCA covariance is symbolized in $\boldsymbol{\Sigma_{w_i,vbpca}}$.

VBPCA covariance matrix is used to calculate the eigenvalues as shown in the following equation:

$$\left|\boldsymbol{\Sigma_{w_i,vbpca}} - \lambda_j \boldsymbol{I}\right| = 0 \qquad (12)$$

Then we get the eigenvalues $\lambda_1,\dots,\lambda_p$ and the eigenvectors $\boldsymbol{a}_1,\dots,\boldsymbol{a}_p$. From the eigenvectors formed, the principal components are obtained using the linear equation given in the following equation:

$$Q_{ij} = \sum_{k=1}^{p} a_{kj}x_{ik}\, , i = 1,2,\dots,n\, ;\, j = 1,2,\dots, \qquad (13)$$

where the principal component of Q_{ij} formed is an orthogonal variable.

The principal components of the VBPCA formed will be regressed using OLS without the intercept of the standard response variables on the *m* principal components as shown in the following equation:

$$y_i = f(Q_1, Q_2, \dots, Q_m) \qquad (14)$$

then obtained the estimated regression coefficient as $\hat{\gamma}_1, \hat{\gamma}_2, \dots, \hat{\gamma}_m$.

From the principal components of the VBPCA formed, the regression coefficients of the linear model based on the principal component regression are obtained which are presented in the following equation.

$$\hat{\beta}_{VBPCR_j} = \frac{s_y}{s_{x_j}}\left[\sum_{t=1}^{m} \boldsymbol{W}_{jt}\hat{\gamma}_t\right] j = 1,2,\dots,p \qquad (15)$$

where $\boldsymbol{W}_{jt}$ is a eigenvectors matrix, then the intercept can be obtained using the following equations:

$$\hat{\beta}_{VBPCR_0} = \bar{y} - \sum_{j=1}^{p} \hat{\beta}_{VBPCR_j}\bar{x}_j \qquad (16)$$

where $s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}$ and $s_{x_j} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)^2}{n-1}}$

2.7 Model Evaluation Criteria

The criteria for evaluating the model in this study used the standard error of regression parameter, value of cross validation (Q^2), adjusted R^2, and MAPE. Standard error of regression parameter estimator can be used to see the effect of multicollinearity. Cross validation is a model selection tool to determine the strength of the model, especially the principal components in PCA [33], adjusted R^2 is used to see the goodness of the regression model formed, while MAPE is used to see the accuracy of the VBPCA method in handling missing data.

2.5.1. Standard Error Value of Regression Parameter Estimation

The standard error of the regression parameter estimator can be used to see the effect of multicollinearity. Multicollinearity on the predictor variables will cause the variance of the regression parameters to be very large so that the standard error of the regression parameter coefficients is also very large. The standard error of the regression parameter estimator formula is given in the following formula [34]:

$$S_e(\hat{\beta}) = \sqrt{\sigma_\varepsilon^2}(\boldsymbol{X'X})^{-1} \qquad (17)$$

where σ_ε^2 is mean square error of regression model.

2.5.2. Cross Validation (Q^2)

Determining the optimal number of principal components that includes relevant information by reducing the presence of noise, one of which can be done by doing cross validation. Cross validation is the most common procedure for determining the number of latent variables used in the model based on the difference between the observed value vector, observed value vector, and predicted value vector [35]. This measure can estimate the level structure of the data set can be optimal in choosing the number of components. The maximum value of cross validation is 1, which means that all variance can be represented in predicting $Y = \hat{Y}$. The formulas for cross validation are given in the following equation [11]:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{d}\left(y_{ij}-\hat{y}_{ij}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{d}\left(y_{ij}\right)^2} \qquad (18)$$

where $\hat{y}_{ij}$ is a predictive value of response variable.

2.5.3. Mean Absolute Percentage Error (MAPE)

In making estimates, there is always a difference between the estimated value and the actual value. This difference in value is called the error value which can be calculated by MAPE [36]. MAPE is a measure to calculate the difference between actual observations and data from forecasts or estimates that are absolute. MAPE is one of the most widely used measures of forecast accuracy due to its superior scale-independence and interpretability [37]. MAPE is considered a good measure of accuracy because MAPE does not depend on the magnitude of the variable to be predicted [38]. A method can be said to have good performance in estimating when it produces a MAPE value of less than 10 [39]. The smaller the resulting MAPE value, the better the method in estimating and forecasting. The MAPE formula is given in the following equation:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{A_i-F_i}{A_i}\right| \quad (19)$$

where A_t is the actual value and F_t is the estimated value.

2.5.4. Adjusted R^2

In general, adjusted R^2 uses a small penalty to add more variables in the model which makes the difference between adjusted R^2 and R^2. In other words, adjusted R^2 penalizes the loss of degrees of freedom resulting from adding predictor variables to the model [40]. Adjusted R^2 is intended to replace the bias estimator with an unbiased component. This leads to the formulation of the adjusted R^2 called the Ezekiel estimator in the statistical literature which is given in the following equation [38]:

$$\text{Ezekiel: } \hat{\rho}_E^2(R^2) = 1 - \frac{N-1}{N-p-1}(1-R^2) \quad (20)$$

where N is the number of observational samples and p is the number of predictor variables. Adjusted R^2 is worth to 0 to 1, the closer to 1, the better the model formed and the variance of predictor variables can describe the variance of response variables very well.

3. RESEARCH METHODS

The method used in this study is a combination of VBPCA with multiple linear regression analysis. The principal component analysis formed from the VBPCA method will be regressed with the response variable using OLS. The VBPCA will discuss the proportion of missing values that has been simulated. The data used in this study is simulation data which is generated from secondary data. The secondary data used is Poverty Depth Index data from three provinces in Indonesia, totaling 100 observations.

There are nine predictor variables and one response variable in secondary data which are used to generate simulation data. The secondary data becomes the initial data of the study, namely complete data. Predictor variables in simulation data will be generated from predictor variables in secondary data by contaminating the collinearity of each predictor variable. Then do the generation of new response variables obtained from multiple linear regression of response variables on complete data and new predictor variables that have been previously generated and contaminated with collinearity. New predictor variables and new response variables have been formed into initial simulation data which will then be simulated with missing data with the percentage of missing data being 1%, 5%, and 10% of the total complete data. There are three final simulation data that have been contaminated with collinearity and there are missing values.

Missing data simulation is carried out using the MCAR mechanism. To obtain data contamination with collinearity on the predictor variable, X_{ik} will be generated using a Monte Carlo simulation with the following equation [41]:

$$X_{ij} = (1-\rho^2)^{\frac{1}{2}}\, x_{ij} + \rho\, x_{ij} \quad (21)$$

where $i = 1,2,\dots,n$ and $j = 1,2,\dots,k$. Then x_{ij} is secondary data and is determined so that the correlation between predictor variables is formulated with ρ^2. The value of ρ is determined at 0.99 which indicates that the variables are highly correlated with each other.

The steps of analysis with simulation data in this study are as follows:

1. Generating simulation data in the form of a matrix containing predictor variables and response variables. Simulation data has been contaminated with collinearity and has missing values on the predictor variables.
2. Testing the assumption of the MCAR data mechanism with the Little's MCAR test on the predictor variables of the simulation data.
3. Testing the assumption of non-multicollinearity with VIF value on the predictor variables for secondary and simulation data.
4. Standardize data for predictor and response variables.

5. Perform principal component analysis using the classical PCA and VBPCA method on the predictor variables of the secondary and simulation data and obtain the principal components.
6. Perform principal component regression with the principal components obtained in step 4. The regression model formed is still in the form regression with the principal component coefficients.
7. Perform back-transformation and substitution of principal components into principal component regression model so that the final regression parameter estimator is obtained.
8. Evaluate the model using cross validation, standard error of regression estimator, MAPE, and adjusted coefficient of determination.
9. Conclusion and suggestion.

4. RESULTS AND DISCUSSION

In this study, the classical PCR method will be compared with the PCR of the VBPCA approach on secondary data and simulation data. The secondary data in this study has complete observations that are different from the simulation data that has been scripted to have missing values. Secondary data in this study was included in the comparison with the aim of seeing how effective the PCR method with VBPCA approach is in overcoming overfitting which is the weakness of several combined methods when faced with different data sets.

4.1 Simulation Scenario

The simulation data used is obtained from the results of real (secondary) data generated in various levels of the proportion of missing data with multicollinearity. The proportion of missing data in the scenario in the predictor variable (X) ant the response variable (Y) has complete observations. The scenario to determine the effect of the proportion of missing data on the predictor variable was designed with three levels of missing data proportion, namely 1%, 5%, 10% of the total value of observations on the predictor variable. The scenario of the simulation data is given in Table 1.

Table 1: Simulation Data Scenario

Total Observations	Collinearity Rate	Proportion of Missing Data	Amount of Missing Data
$n \times p$ $= 100 \times 9$ $= 900$	$\rho = 0.99$	1%	9
		5%	45
		10%	90

Based on the scenario of the simulation data presented in Table 1, three data combinations will be tabulated in the form of matrix where each data has a combination of collinearity and missing data on the predictor variable. In the simulation data, the predictor variables for all simulation data will have 9, 45, and 90 missing values, respectively. The simulated missing data scenario spreads randomly according to the MCAR mechanism. The simulated missing data plot is presented in Figure 1.

Based on figure 1, the missing observations that have been simulated on the predictor variables for all simulation data visually have a random and unstructured pattern. The random pattern has been designed according to the MCAR missing data mechanism. The simulated missing data pattern is also randomly distributed for each predictor variable and has the appropriate amount, so that the simulation results meet the designed simulation scenario. To the further prove that the simulated missing data pattern has met the MCAR assumptions.

4.2 Missing Data Mechanism Test

Testing the assumption of missing data mechanism used Little's MCAR test. The assumption test is carried out for all simulation data with α = 5%. The test results are presented in Table 2.

Table 2: Little's MCAR Test Results

Proportion of Missing Data	Little's MCAR (d^2)	p-value
1%	60.3	0.258
5%	137	0.692
10%	226	0.888

Based on the test results in Table 2, it can be seen that the p-value of the Little's MCAR test statistic for the entire proportion of missing data is greater than α = 5% so that according to the hypothesis, we can accept the null hypothesis. Thus, it can be concluded that all simulation data meet the assumption of the MCAR mechanism. This means that the pattern of missing data that is formed occurs randomly and not systematically.

4.3 Non-Multicollinearity Assumption Test

Testing the non-multicollinearity assumption on the predictor variable used the VIF value. Testing the non-multicollinearity assumption is aimed at seeing whether the collinearity simulation carried out is in accordance with the expected scenario. To get the VIF value, it's necessary to do multiple linear regression modeling on the simulation data. VIF values for all simulation data and secondary data are presented in Table 3.

Table 3: Simulation Data Scenario

Variable	Secondary Data	Simulation Data		
		Missing Data 1%	Missing Data 5%	Missing Data 10%
X_1	26.49	4,101.16	5,631.13	18,927.67
X_2	72.50	158.98	185.66	287.41
X_3	65.61	67.77	67.83	267.82
X_4	463.43	3,858.75	5,220.12	16,885.90
X_5	2.81	2.91	3.86	6.39
X_6	21.27	3,058.01	3,796.51	8,937.68
X_7	2.15	3.91	4.49	14.610
X_8	3.59	2,779.43	3,428.46	7,699.06
X_9	2.54	138.52	154.35	354.61

Based on the VIF values in Table 3, it can be seen that all simulation data and secondary data have predictor variables that have a VIF value of more than 10. This indicates that the entire proportion of missing data in the simulation data has been contaminated with collinearity. So that the collinearity simulation carried out is in accordance with the expected scenario.

4.4 Estimation of Classical Principal Component Regression Parameters

The first step in classical PCR is to determine the covariance matrix. The covariance matrix is used if the data units are the same, in accordance with the secondary and simulation data in this study which had been standardized prior to PCA. Then, the covariance matrix is used to get the eigenvalues and eigenvectors. The eigenvalues can be visualized using a scree plot. The scree plot aims to make it easier to see the pattern of eigenvalues formed for each principal component. The scree plot of the eigenvalues for secondary and simulation data is presented in Figure 2.

In this study, all principal components were selected for PCA modelling, so that all information on principal components could be summarized in the model. The eigenvalues formed will be used to obtain the eigenvectors. Then the eigenvectors or called loadings values will be formed into a principal component model. There are 9 principal component models for each secondary and all simulation data.

Based on the classical principal component model that was formed, then the PCR analysis was carried out by regressing the principal component Q formed from the principal component model to the response variable. The principal component Q is formed from entering all components of the predictor variable into the principal component model that is formed. The model is still in regression form with principal component coefficients, so it needs to be substituted back into the PCR model. In addition, the model coefficients are still in the form of results from standardized data, so they need to be returned to the initial observation unit. The results of the estimation of the final parameter coefficients of the classical PCR model are presented in Table 4. Based on the results of the estimation of classical PCR parameters formed in Table 4, the final classical PCR model for secondary and simulation data is given as follows.

Classical PCR Model for Secondary Data

$$\hat{Y} = 7.2000 + 0.1152X_1 - 0.0879X_2 + 0.0003X_3 \\ -0.2587X_4 - 0.2568X_6 + 0.0283X_8 - 0.0009X_9$$

Classical PCR Model for Simulation Data (Missing Data 1%)

$$\hat{Y} = 7.8753 - 0.0835X_1 - 0.0445X_2 + 0.0003X_3 \\ -0.0835X_4 - 0.4789X_6 + 0.1685X_8 - 0.0414X_9$$

Classical PCR Model for Simulation Data (Missing Data 5%)

$$\hat{Y} = 11.5823 + 1.0830X_1 + 0.2092X_2 + 0.0028X_3 \\ -1.0603X_4 + 1.7336X_6 - 1.6096X_8 - 0.2519X_9$$

Classical PCR Model for Simulation Data (Missing Data 10%)

$$\hat{Y} = 1.1175 - 2.3236X_1 - 1.1993X_2 - 0.0056X_3 \\ +2.2375X_4 - 8.6233X_6 + 7.9150X_8 + 0.4790X_9$$

4.5 Estimation of Variational Bayesian Principal Component Regression Parameters

The first step in the PCR with VBPCA method is to analyze the principal components of the VBCPA, then do a regression using the principal components generated by the VBPCA method. The principal component of the VBPCA method uses a prior distribution which is used to estimate missing data. The variables used in the prior are $\gamma_{\mu_0}, \bar{\boldsymbol{\mu}}_0, \gamma_{\tau_0}$ and $\bar{\boldsymbol{\tau}}_0$, where all of these variables are hyperparameters that define the prior. The value of the hyperparameter has been set according to the non-informative prior, namely $\gamma_{\mu_0} = \gamma_{\tau_0} = 10^{-10}$, $\bar{\boldsymbol{\mu}}_0 = 0$ and $\bar{\boldsymbol{\tau}}_0 = 1$. So that the posterior distribution is obtained by marginalizing the likelihood function. Based on these priors and posteriors, missing data can be estimated to obtain a complete data set.

In secondary data, all observations are complete or there are no missing data. In the calculation algorithm, the process of estimating missing data using the VBPCA method on secondary data is still running. Complete data acquisition process for missing data estimation using VBPCA, produces the same results as secondary data. So that in secondary data, the classical PCR process and the PCR with VBPCA method will produce the same results and model.

In the simulation data, there are missing data so that the missing data will be estimated using priors and posteriors, then the complete data set is obtained. From the complete data set, PCA will be carried out to overcome multicollinearity. Based on the estimation results of missing data, new observations which are the results of the estimation of missing data using the VBPCA method on missing data simulation of 1%, 5%, and 10%, so that a complete data set has been formed entirely. Furthermore, the complete data set was analyzed for principal components to overcome multicollinearity. From the complete data set that has been formed for simulation data, covariance matrix is obtained. Based on the covariance matrix, then the eigenvalues can be formed. In all simulation data for the PCR with VBPCA approach, the eigenvalues formed are visualized using a scree plot. The scree plot for the eigenvalues for simulation data is presented in Figure 3.

Based on the eigenvalues formed, the eigenvectors are obtained which then become the loadings values for the loading values for the principal component model. After obtaining the model for the principal components of the VBPCA method, a PCR analysis was carried out, namely regressing the principal components of the VBPCA method formed on the response variable from the simulation data using OLS method. The model that is formed is still in the form of regression with the principal component coefficients so that it needs to be substituted back into the PCR model using predictor variables. The results of the estimation of the final parameter coefficients of the classical PCR with VBPCA approach are presented in Table 5. Based on the results of the estimation of the PCR with VBPCA method formed in Table 5, the final PCR model for secondary and all simulation data is given in the following model.

PCR Model with VBPCA Approach for Secondary Data

$$\hat{Y} = 7.2000 + 0.1152X_1 - 0.0879X_2 + 0.0003X_3$$
$$-0.2587X_4 - 0.2568X_6 + 0.0283X_8 - 0.0009X_9$$

PCR Model with VBPCA Approach for Simulation Data (Missing Data 1%)

$$\hat{Y}_i = 7.3400 - 0.2443X_{1i} - 0.0156X_{2i} + 0.0000X_{3i}$$
$$-0.1354X_{4i} - 0.0021X_{5i} - 0.5817X_{6i}$$
$$-0.0024X_{7i} + 0.8475X_{8i} - 0.0560X_{9i}$$

PCR Model with VBPCA Approach for Simulation Data (Missing Data 5%)

$$\hat{Y}_i = 8.9260 - 0.2443X_{1i} - 0.1301 + 0.0001X_{3i}$$
$$-0.2500X_{4i} - 0.0093X_{5i} - 0.9056X_{6i}$$
$$-0.0057X_{7i} + 1.8233X_{8i} - 0.1059X_{9i}$$

PCR Model with VBPCA Approach for Simulation Data (Missing Data 10%)

$$\hat{Y}_i = 8.2757 - 0.2678X_{1i} - 0.0580X_{2i} + 0.0009X_{3i}$$
$$-0.1159X_{4i} - 0.0011X_{5i} + 0.2545X_{6i}$$
$$-0.0004X_{7i} - 0.0056X_{8i} - 0.1000X_{9i}$$

4.6 Model Evaluation

4.6.1. Standard Error Value of Regression Parameter Estimation

The standard error of regression parameter estimator for secondary data and simulation data uses multiple linear regression analysis. The results of the standard error of multiple linear regression will be a comparison for the classical PCR and the PCR with VBPCA method, which then the classical PCR will be compared with the PCR with VBPCA method. The results of the comparison of the standard error coefficients of the regression parameters are presented in Table 6. The comparison of standard error values of the regression parameter estimators in Table 6 is visualized in graphic form in Figure 4.

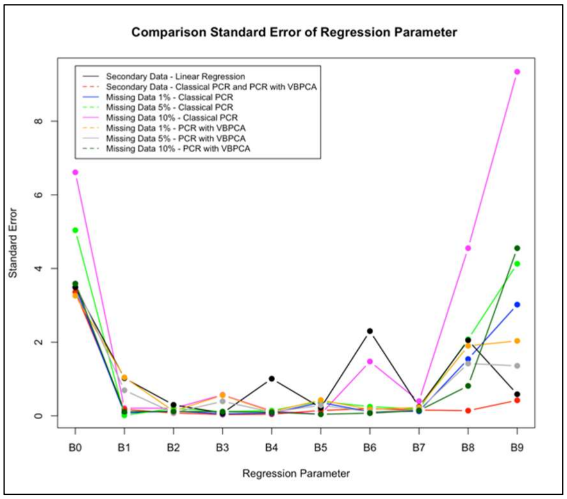


Figure 4: Standard Error of Regression Parameter Estimator Comparison

Based on the standard error values of the regression parameter estimators presented in Table 6 and Figure 4, in the secondary data it can be seen that the standard error of the multiple linear regression parameter estimator is greater than the standard error of the classical PCR and the PCR with VBPCA method. This is because in multiple linear regression, the multicollinearity contained in the predictor variables has not been resolved, thus making the variance and standard error of the regression parameter estimator larger. The greater variance and standard error will cause the confidence interval of the regression parameter estimator to be greater so that the opportunity to accept the null hypothesis will be greater. Meanwhile, the classical PCR method and the PCR with VBPCA method for secondary data have a smaller standard error value for the regression parameter. This is because multicollinearity of the predictor variables has been resolved. The results of the estimation of the standard error value of the regression parameters for the classical PCR and the PCR with VBPCA method have the same value, because the secondary data has complete observations so that the process of estimating missing observations with the VBPCA method is not used.

In the simulation data, the classical PCR and the PCR with VBPCA method have differences in the standard error value of the regression parameter estimator. The standard error value of the regression parameter estimator in PCR with VBPCA for all simulation data is greater than the classical PCR. The PCR with VBPCA method overcomes missing data, so that complete observations will be obtained that are in accordance with the secondary data. When compared with classical PCR, the classical PCR method does not address missing data so that the PCR process will use available information without using information from missing observations. In addition, the greater the proportion of missing data, the greater the standard error of the PCR parameter estimator. This is because the missing data will cause the information in the data to be reduced and make the variance of the data bigger which causes the analysis results to be less valid.

4.6.2. Cross Validation (Q^2)

The optimal principal component in the classical PCR as well as the VBPCA approach will be indicated by the value of Q^2. In this study, the value of Q^2 of the secondary data will be compared with the simulation data for all methods, namely classical PCR and PCR with VBPCA approach. The value of Q^2 from the analysis is presented in Table 7. The comparison of cross validation of the principal components in Table 7 is visualized in graphical form in Figure 5.

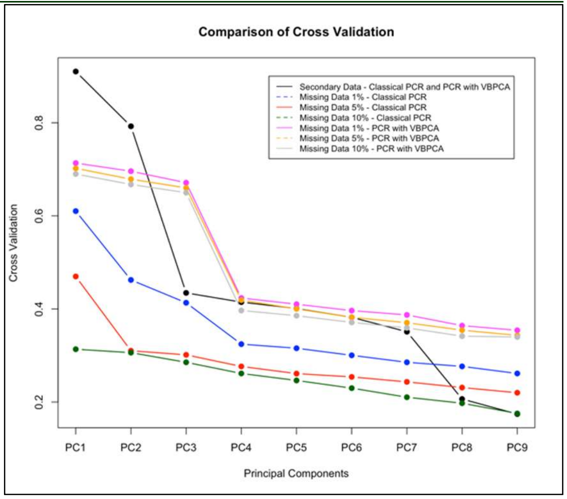


Figure 5: Principal Component Cross Validation

Based on the Q^2 values in Table 7 and Figure 5, it can be seen that in secondary data, the classical PCR method and PCR with VBPCA have the same Q^2 value caused by a complete set of observations. In addition, the secondary data has a value of Q^2 which tends to be higher than other methods because there are no missing data, so that all information formed by the principal components can explain the data optimally. Meanwhile, in the simulation data, the PCR method with VBPCA has a higher Q^2 value than the classical PCR for all principal components. This is because in the VBPCA method, missing data have been estimated so that the observations used to form the principal component come from a complete data set so that the information covered by the principal component to explain the variance of the data is higher than the PCR. For all methods on the simulation data and secondary data, the first three components have a higher Q^2 value than the 4-th to 9-th components, it shows that the initial three principal components are sufficient to explain the variance and provide sufficient information from the data. In addition, the proportion of missing data also affects how the principal components explain the variability of the data. The higher the proportion of missing data, the lower the Q^2 value of the principal component and vice versa.

4.6.3. Mean Absolute Percentage Error (MAPE)

In this study, the MAPE value was used to see the accuracy of the estimation results of missing data by the VBPCA method. The VBPCA method will be said to have good performance in overcoming missing data if the resulting MAPE value is less than 10%. The smaller the resulting

MAPE value, the better the VBPCA method in overcoming the missing data values. MAPE calculations are only used on simulation data because secondary data has complete data. The results of the MAPE values for the VBPCA method on the entire proportion of missing data are given in Table 8.

Table 8: Comparison of MAPE Values

Simulation Data	MAPE
Missing Data 1%	1.53%
Missing Data 5%	6.04%
Missing Data 10%	10.13%

Based on the MAPE calculation results for all missing simulation data in Table 8, it can be seen that for 1% and 5% missing data, the resulting MAPE value for the VBPCA method is less than 10%. Meanwhile for 10% missing data, the MAPE value of the VBPCA method is 10.13% which is worth more than 10%. This shows that the VBPCA method can handle missing data up to less than 10% of missing observations from all complete data.

4.6.4. Adjusted $\mathbf{R^2}$

The evaluation of the model to see the goodness of the PCR model formed in this study used adjusted R^2. The value of adjusted R^2 will show how much the variance of the response variables can be explained by the variance of the predictor variables. The results of the calculation of the adjusted R^2 for all methods are presented in Table 9.

Table 9: Comparison of Adjusted R^2

Data	Method	
	Classical PCR	PCR with VBPCA
Secondary Data	0,4894	0,4894
Missing Data 1%	0.3949	0.4048
Missing Data 5%	0.3855	0.4043
Missing Data 10%	0.3765	0.4028

The comparison of the *Adjusted* R^2 in Table 9 is visualized in graphic form in Figure 6.

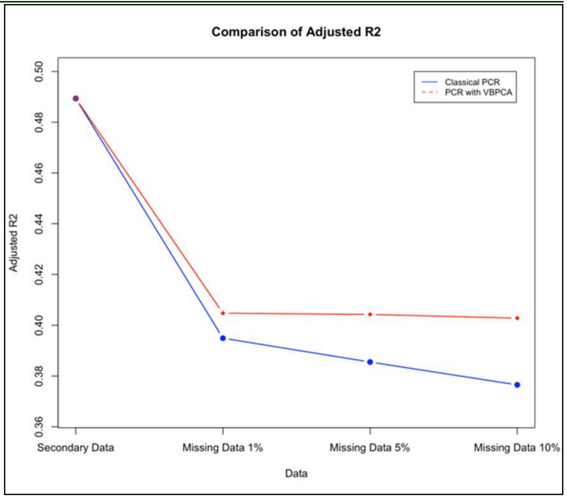


Figure 6: Comparison of Adjusted R^2

Based on the calculation results of the adjusted R^2 value in Table 9 and Figure 6, it can be seen that at all levels the proportion of missing data simulation, the value of the adjusted R^2 for the PCR with VBPCA method is higher than the classical PCR method. The higher the proportion of missing observations in a data, the smaller the value of the adjusted R^2 and vice versa. The number of missing data is the cause of the lower value of the adjusted R^2 formed. In classical PCR, missing data are not estimated so that the principal component formed stores observational information without the information contained in missing data, so that in the PCR process the variance of predictor variables is also not optimal in explaining the variance of response variables. So that the PCR with the VBPCA method will be better in overcoming this problem so that the variance of predictor variables will be more optimal in explaining the variance of response variables. In addition, in the comparison graph of the adjusted R^2, it can be seen that the value of the adjusted R^2 for the classical PCR method and PCR with VBPCA has the same value in the secondary data. The VBPCA method is used to overcome missing data, but when faced with complete data, the VBPCA method produces the same coefficient of determination as the classical PCA. This can be an indicator that proves that the VBPCA method can overcome the problem of overfitting, where this problem often occurs when the method is faced with data that has been scripted according to the needs of the method.

5. CONCLUSIONS AND SUGGESTIONS

Based on the applied studies and simulation studies conducted in this research, the following conclusions can be drawn:

1. The estimation of the PCR parameters using the VBPCA method using OLS in this study shows that the results of the regression parameter estimation have a constant value at the proportion of missing data up to 5%. The analysis shows that the greater the proportion of missing data, the more inconsistent the estimation results of the regression parameters and the larger the standard error of the regression parameters.
2. Based on the evaluation of the model for classical PCR and PCR with VBPCA, it is concluded that:
   a. PCR with VBCPA is better than classical PCR and multiple linear regression in overcoming multicollinearity which is indicated by a smaller standard error value of the regression parameter.
   b. The MAPE value for the proportion of missing data of 1% and 5% is less than 10%, while for the proportion of missing data 10%, the resulting MAPE value is 10.13%, so the VBPCA method can handle missing data up to less than 10%.
   c. The level of the proportion of missing data affects the measure of the goodness of the model. The higher the proportion of missing data, the less optimal the resulting model is indicated by the larger MAPE value, the smaller the value of cross validation and adjusted R^2. The proportion of missing data causes the information obtained from the predictor variables to be less than optimal.
   d. PCR model with VBPCA can overcome overfitting indicated by the adjusted R^2 value which is the same as the classical PCR method on secondary data that has complete data.

The strength of this study is that it is able to overcome missing data up to 10%. In addition, the strength if this study us that the method formed can overcome overfitting, which often occurs in a combination of two or methods. The weakness in this research is the main component analysis, namely the missing data. In addition, one of the criteria for the goodness of the model is the coefficient of determination in this study is small. This can be due to the selection of the independent variables that are not suitable.

Suggestions that can be given by researchers for further research based on the conclusions are that this study has not examined the effect of the level of multicollinearity with missing data on PCR with VBPCA, so that in future research, the level of multicollinearity can be used to determine the effect of the level of multicollinearity with missing data. In addition, this study uses a large sample and has not studied the use of a small sample in PCR with VBPCA, so that in future research, a small sample can be used to determine the effectiveness of the PCR method with VBPCA in overcoming multicollinearity and missing data in small samples.

Research contributions are:

- NABILA AZARIN BALQIS as the main author with the task of being in charge of the paper, the process of improving the paper, revising the paper.
- SUCI ASTUTIK is the second author to have contributed in the literature suggestions
- SOLIMUN is the third author who has contributed or is responsible for research methodology
- NURJANNAH is the fourth author who has contributed to the process of collecting data and designing the preparation of research results
- HENNY PRAMOEDYO is the fifth author who has contributed to the process of data analysis and interpretation of research results.
-

REFERENCES:

[1] H. Kim and H. Y. Jung, “Ridge Fuzzy Regression Modelling for Solving Multicollinearity”, *Mathematics,* Vol. 8, No. 9, 2020, doi: 10.3390/math8091572.

[2] K. Ayinde, A. F. Lukman, O. O. Alabi, and H. A. Bello, “A New Approach of Principal Component Regression Estimator with Applications to Collinear Data”, *International Journal of Engineering Research and Technology,* Vol. 13, No. 7, 2020, pp. 1616-1622, doi: 10.37624/IJERT/13.7.2020.1616-1622.

[3] H. Kang, “The Prevention and Handling of the Missing Data”, *Korean Journal of Anesthesiology,* Vol. 64, No. 5, 2013, pp. 402-406, doi: 10.4097/kjae.2013.64.5.402.

[4] R. H. H. Groenwold and O. M. Dekkers, “Missing Data: The Impact of what is not there”, *European Journal of Endocrinology,* Vol. 183, 2020, doi: 10.1530/EJE-20-0732.

[5] A. Agarwal, D. Shah, D. Shen, and D. Song, “On Robustness of Principal Component Regression”, *Journal of the American Statistical Association,* Vol. 00, No. 0, 2021, pp. 1731-1745, doi: 10.1080/01621459.2021.1928513

[6] J. R. Carpenter and M. Smuk, “Missing data: A Statistical Framework for Practice”, *Biometric Journal,* Vol. 3, 2021, pp. 915-947, doi:

10.1002/bimj.202000196.

[7] C. M. Bishop, “Variational Principal Component”, *Ninth International Conference on Artificial Neural Networks*, ICANN, IEE. Vol I, 1999, pp. 509- 514.

[8] S. Oba, M. Sato, and S. Ishii, “Prior Hyperparameters in Bayesian PCA”, *Joint International Conference ICANN/ICONIP*, LNCS 2714, 2003, pp. 271-279.

[9] W. Stacklies, H. Redestig, M. Scholz, D. Walther, and J. Selbig, “A Bioconductor Package Providing PCA Methos for Incomplete Data”, *Bioinformatics,* Vol. 23, No. 9, 2007, pp. 1164-1167.

[10] W. Li, W. Jiang, Z. Li, H. Chen, Q. Chen, J. Wang, and G. Zhu, “Extracting Common Mode Errors of Regional GNSS Position Time Series in the Presence of Missing Data by Variational Bayesian Principal Component Analysis”, *Sensors,* Vol. 20, No. 8, 2020, doi: 10.3390/s20082298.

[11] R. Yordani, “Penerapan Model Inferensi Bayesian dengan Variational Bayesian Principal Component Analysis (VBPCA) dalam Mengatasi Missing Data Analisis Komponen Utama”, *Jurnal Aplikasi Statistika & Komputasi Statistik,* Vol. 7, No. 2, 2015.

[12] Taufiq, A., Astuti, A. B., & Fernandes, A. A. R. “Geographically Weighted Regression in Cox Survival Analysis for Weibull Distributed Data with Bayesian Approach.” *In IOP Conference Series: Materials Science and Engineering*, 2019, Vol. 546, No. 5, p. 052078. IOP Publishing.

[13] Fernandes, A. A. R. & Solimun. S. “Moderating effects orientation and innovation strategy on the effect of uncertainty on the performance of business environment.” *International Journal of Law and Management*, 2017, *59*(6), 1211-1219.

[14] Raharjo, K., Nurjannah, N., Solimun, S., & Fernandes, A. A. R. “The influence of organizational culture and job design on job commitment and human resource performance.” *Journal of Organizational Change Management,* 2018.

[15] J. W. Graham, “Missing Data Analysis: Making it Work in the Real World”, *Annu Rev Psychol*, Vol. 60, pp. 549-576, 2009.

[16] D. A. Bennet, “How Can I Deal with Missing Data in My Study?”, *Aust N Z J Public Health*, Vol. 25, No. 5, pp. 464-469, 2001.

[17] K. J. Lee, K. M. Tilling, R. P. Cornish, R. J. A. Little, M. L. Bell, E. Goetghebeur, J. W. Hogan, and J. R. Carpenter, “Framework for the Treatment and Reporting of Missing Data in Observational Studies: The Treatment and Reporting of Missing Data in Observational Studies Framework”, *Journal of Clinical Epidemology*, Vol. 134, 2021, pp. 79-88, doi: 10.1016/j.jclinepi.2021.01.008.

[18] C. G. Marcelino, G. M. C. Leite, P. Celes, and C. E. Pedreira, “Missing Data Analysis in Regression”, *Applied Artificial Intelligence,* Vol. 36, No. 1, 2022, doi: 10.1080/08839514.2022.2032925.

[19] T. Tsiampalis and D. B. Panagiotakos, “Missing-data Analysis: Socio-demographic, Clinical and Lifestyle Determinants of Low Response Rate on Self-reported Psychological and Nutrition Related Multi-Item Instruments in the Context of the ATTICA Epidemiological Study”, *BMC Medical Research Methodology*, Vol. 20, No. 148, 2020, doi: 10.1186/s-12874-020-01098-3.

[20] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, “A Surver on Missing Data in Machine Learning”, *Journal of Big Data*, Vol. 8, No. 140, 2021, doi: 10.1186/s40537-021-00516-9.

[21] B. G. Tabachnick and L. S. Fidell, “Using Multivariate Statistics”, Allyn & Bacon, Needham Heights, 2012.

[22] R. J. A. Little and D. B. Rubin. “Statistical Analysis with Missing Data”, John Wiley and Sons, Hoboken, 1987.

[23] A. Z. Alruhaymi and C. J. Kim, “Study on the Missing Data Mechanisms and Imputation Methods”, *Open Journal of Statistics,* Vol. 11, 2021, pp. 477-492, doi: 10.4236/ojs.2021.114030.

[24] R. J. A. Little, “A Test of Missing Completely at Random for Multivariate data With Missing Values”, *Journal of the American Statistical Association*, Vol. 83, 1988, pp. 1198-1202.

[25] O. O. Alabi, K. Ayinde, O. E. Babalola, H. A. Bello, and E. C. Okon, “Effects of Multicollinearity on Type I Error of Some Methods of Detecting Heteroscedasticity in Linear Regression Model”, *Open Journal of Statistics,* Vol. 10, 2020, pp. 664-677, doi: 10.4236/ojs.2020.104041.

[26] A. U. Ahmad, U. V. Balakrishnan, and P. S. Jha, “A Study of Multicollinearity Detection and Rectification under Missing Values”, *Turkish Journal of Computer and Mathematics Education,* Vol. 12, No. 1S, 2021, pp. 399-418.

[27] B. L. Bowerman, and R. T. O'Connel, "Linear Statistical Models: An Applied Approach". Cole: Brooks, 1990.

[28] N. R. Draper and H. Smith, "Analisis Regresi Terapan, Edisi Kedua", Jakarta: Gramedia Pustaka Umum, 1992.

[29] D. C. Montgomery and E. A. Peck, "Introduction to Linear Regression Analysis, Second Edition", New York: John Wiley & Sons, Inc, 1992.

[30] Notiragayu dan K. Nisa, "Analisis Regresi Komponen Utama Robust untuk Data Mengandung Pencilan", *Jurnal Sains MIPA*, Vol. 14, No. 1, 2008, pp. 45-50.

[31] M. R. Mahmouedi, M. H. Heydarim S. N. Oasem, and S. S. Band, "Principal Component Analysis to Study the Relations Between the Spread Rates of COVID-19 in High Risks Countries", *Alexandria Engineering Journal*, Vol. 60, 2020, pp. 457-464, doi: 10.1016/j.aej.2020.09.013.

[32] I. T. Jolliffe and J. Cadima, "Principal Component Analysis: A Review and Recent Developments", *Philos. Trans. R. Soc.,* Vol. 374, 2016, doi: 10.1098/rsta.2015.0202.

[33] N. C. Schipper and K. V. Deun, "Model Selection Techniques for Sparse Weight-based Principal Component Analysis", *Journal of Chemometrics,* Vol. 35, 2021, doi: 10.1002/cem.3289.

[34] O. L. O. Astivia and B. D. Zumbo, "Heteroskedasticity in Multiple Regression Analysis: What it is, How to Detect it and How to Solve it with Applications in R and SPSS", *Practical Assessment, Research, and Evaluation,* Vol. 24, No. 1, 2019, doi: 10.7275/q5xr-fr95.

[35] D. N. Rutledge, J. M. Roger, M. and Lesnoff, "Different Methods for Determining the Dimensionality of Multivariate Models", *Frontiers in Analytical Science,* Vol. 1, 2021, doi: 10.3389/frans.2021.754447.

[36] F. Liantoni and A. Agusti, "Forecasting Bitcoin Using Double Exponential Smoothing Method Based on Mean Absolute Percentage Error", *International Journal on Informatics Visualization,* Vol. 4, No. 2, 2020, doi: 10.30630/joiv.4.2.335.

[37] S. Kim and H. Kim, "A New Metric of Absolute Percentage Error for Intermittent Demand Forecasts", *International Journal of Forecasting,* Vol. 32, 2016, pp. 669-679, doi: 10.1016/j.ijforecast.2015.12.003.

[38] M. D. Estrada, M. E. Camarillo, M. Parraguirre, M. E. Castillo, E. Juarez, and M. J. Gomez, "Evaluation of Several Error Measures Applied to the Sales Forecast System of Chemicals Supply Enterprises", *International Journal of Business Administration,* Vol. 11, No. 4, 2020, doi: 10.5430/ijba.v11n4p39.

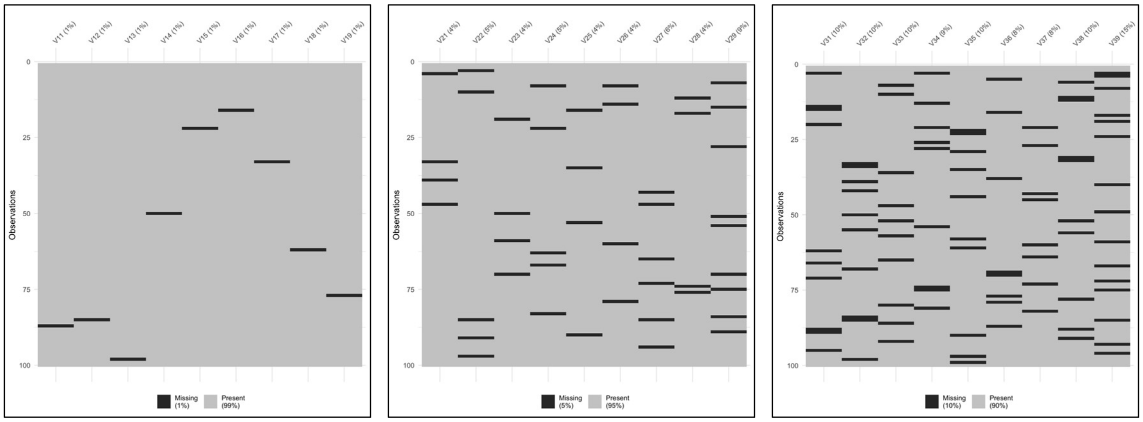
[39] M. Arumsari dan A. T. R. Dani, "Peramalan Dara Runtun Waktu Menggunakan Model Hybrid Time Series Regression - Autoregressive Integrated Moving Average". *Jurnal Siger Matematika,* Vol. 02, No. 01, 2021.

[40] H. Pham, "A New Criterion for Model Selection", *Mathematics,* Vol. 7, No. 1215, 2019, doi: 10.3390/math7121215.

[41] J. Karch, "Improving on Adjusted R-squared", *Collabra: Psychology*, Vol. 6, No. 1, 2020, pp. 45.

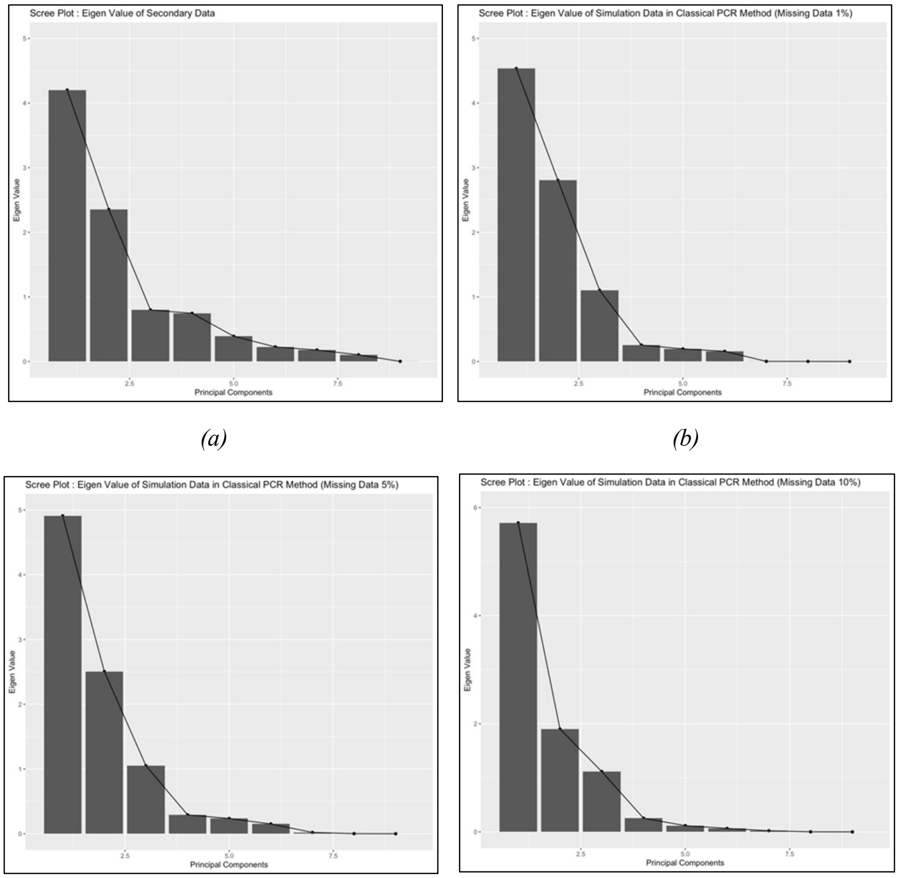
[42] G. C. McDonald and D. I. Galarneau. "A Monte Carlo Evaluation of Some Ridge-Type Estimators", *J. Amer. Statist. Asoc*, Vol. 70, 1975, pp. 407-416

APPENDIX



(a) (b) (c)

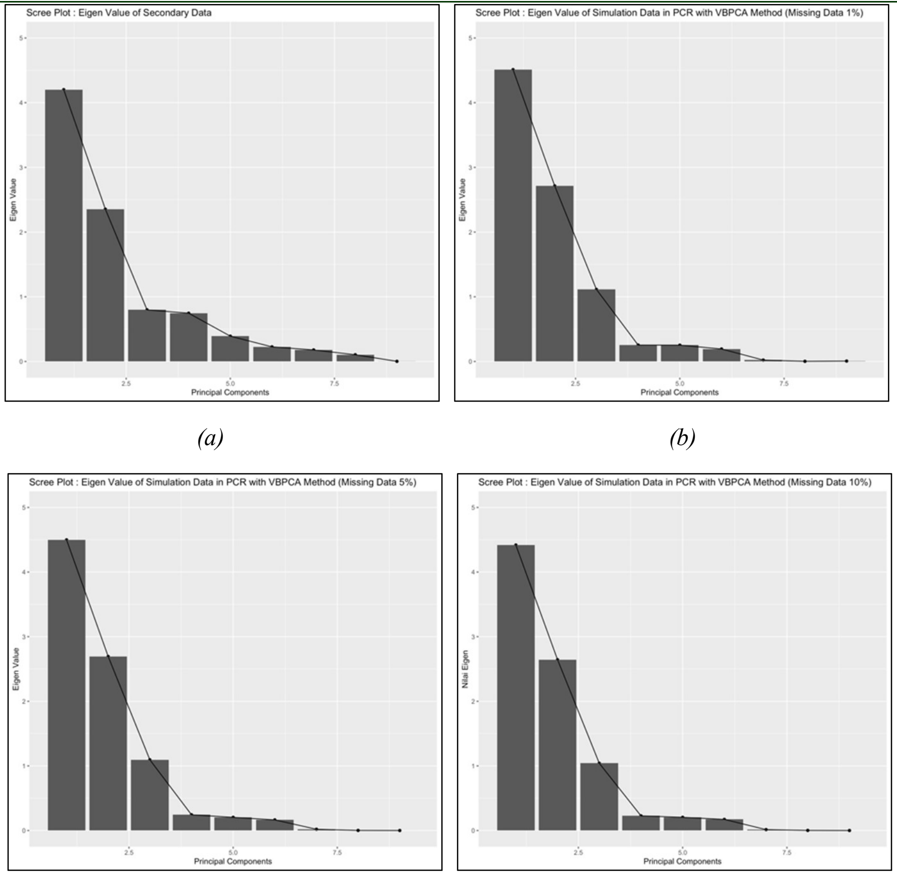
Figure 1: Missing Data Proportion (a) 1%, (b) 5%, (c) 10%



(a) (b)

(c) (d)

Figure 2: Scree Plot of Eigen Values (Classical PCR Method) (a) Secondary Data, (b) Missing Data 1%, (c) Missing Data 5%, (d) Missing Data 10%



(a) *(b)*

(c) *(d)*

Figure 3: Scree Plot of Eigen Values (PCR with VBPCA Method) (a) Secondary Data, (b) Missing Data 1%, (c) Missing Data 5%, (d) Missing Data 10%

Table 4: Parameters Estimate of Classical PCR Method

Parameter	Secondary Data		Missing Data 1%		Missing Data 5%		Missing Data 10%	
	$\hat{\beta}$	$\left(S_e(\hat{\beta})\right)$	$\hat{\beta}$	$\left(S_e(\hat{\beta})\right)$	$\hat{\beta}$	$\left(S_e(\hat{\beta})\right)$	$\hat{\beta}$	$\left(S_e(\hat{\beta})\right)$
β_0	7.2004	3.3634	7.8753	3.4899	11.5823	5.0392	1.1175	6.6114
β_1	0.1152	0.1747	0.0835	0.0761	1.0830	0.1039	-2.3236	0.2077
β_2	-0.0879	0.0767	-0.0445	0.1317	0.2092	0.2009	-1.1993	0.2048
β_3	0.0003	0.0351	0.0003	0.0538	0.0028	0.1151	-0.0056	0.5691
β_4	-0.2587	0.0440	-0.0835	0.0745	-1.0603	0.1428	2.2375	0.1316
β_5	0.0000	0.1468	0.0000	0.3627	0.0000	0.3935	0.0000	0.0470
β_6	-0.2568	0.1941	-0.4789	0.1094	1.7336	0.2491	-8.6233	1.4747
β_7	0.0000	0.1584	0.0000	0.1232	0.0000	0.1842	0.0000	0.3988
β_8	0.0283	0.1405	0.1685	1.5408	-1.6096	2.0777	7.9150	4.5516
β_9	-0.0009	0.4212	-0.0414	3.0195	-0.2519	4.1289	0.4790	9.3413

Source: Secondary and Simulation Data Processed (2022)

Table 5: Parameters Estimate of PCR with VBPCA Method

Parameter	Secondary Data		Missing Data 1%		Missing Data 5%		Missing Data 10%	
	$\hat{\beta}$	$\left(S_e(\hat{\beta})\right)$	$\hat{\beta}$	$\left(S_e(\hat{\beta})\right)$	$\hat{\beta}$	$\left(S_e(\hat{\beta})\right)$	$\hat{\beta}$	$\left(S_e(\hat{\beta})\right)$
β_0	7.2004	3.3634	7.3400	3.2617	8.9260	3.5957	8.2757	3.5895
β_1	0.1152	0.1747	-0.2443	1.0432	-0.4836	0.6947	-0.2678	0.1133
β_2	-0.0879	0.0767	-0.0156	0.1293	-0.1301	0.0830	-0.0580	0.1175
β_3	0.0003	0.0351	0.0000	0.5585	0.0001	0.3941	0.0009	0.1134
β_4	-0.2587	0.0440	-0.1354	0.1126	-0.2500	0.1002	-0.1159	0.0985
β_5	0.0000	0.1468	-0.0021	0.4277	-0.0093	0.2875	-0.0011	0.0423
β_6	-0.2568	0.1941	-0.5817	0.1708	-0.9056	0.1096	0.2545	0.0723
β_7	0.0000	0.1584	-0.0024	0.2311	-0.0057	0.1635	-0.0004	0.1444
β_8	0.0283	0.1405	0.8475	1.9007	1.8233	1.4159	-0.0056	0.8134
β_9	-0.0009	0.4212	-0.0560	2.0343	-0.1059	1.3582	-0.1000	4.5516

Source: Secondary and Simulation Data Processed (2022)

Table 6: Standard Error of Regression Parameters Estimator

Parameter	Secondary Data			Simulation Data					
	Linear Regression	Classical PCR	PCR with VBPCA	Classical PCR			PCR with VBPCA		
				1%	5%	10%	1%	5%	10%
β_0	3.4900	3.3634	3.3634	3.4899	5.0392	6.6114	3.2617	3.5957	3.5895
β_1	1.0230	0.1747	0.1747	0.0761	0.1039	0.2077	1.0432	0.6947	0.1133
β_2	0.2981	0.0767	0.0767	0.1317	0.2009	0.2048	0.1293	0.0830	0.1175
β_3	0.0748	0.0351	0.0351	0.0538	0.1151	0.5691	0.5585	0.3941	0.1134
β_4	1.0080	0.0440	0.0440	0.0745	0.1428	0.1316	0.1126	0.1002	0.0985
β_5	0.2196	0.1468	0.1468	0.3627	0.3935	0.0470	0.4277	0.2875	0.0423
β_6	2.3010	0.1941	0.1941	0.1094	0.2491	1.4747	0.1708	0.1096	0.0723
β_7	0.2458	0.1584	0.1584	0.1232	0.1842	0.3988	0.2311	0.1635	0.1444
β_8	2.0520	0.1405	0.1405	1.5408	2.0777	4.5516	1.9007	1.4159	0.8134
β_9	0.5861	0.4212	0.4212	3.0195	4.1289	9.3413	2.0343	1.3582	4.5516

Source: Secondary and Simulation Data Processed (2022)

Table 7: Cross Validation (R^2) *Value*

Principal Component	Method							
	Classical PCR				PCR with VBPCA			
	Secondary Data	Missing Data 1%	Missing Data 5%	Missing Data 10%	Secondary Data	Missing Data 1%	Missing Data 5%	Missing Data 10%
Q_1	0.9101	0.6102	0.4698	0.3135	0.9101	0.7134	0.7021	0.6899
Q_2	0.7942	0.4623	0.3102	0.3061	0.7942	0.6959	0.6788	0.6675
Q_3	0.4365	0.4133	0.3014	0.2854	0.4365	0.6712	0.6601	0.6498
Q_4	0.4147	0.3244	0.2765	0.2614	0.4147	0.4235	0.4189	0.3967
Q_5	0.4009	0.3156	0.2611	0.2464	0.4009	0.4102	0.4001	0.3855
Q_6	0.3820	0.3004	0.2540	0.2299	0.3820	0.3964	0.3821	0.3711
Q_7	0.3511	0.2854	0.2433	0.2104	0.3511	0.3872	0.3704	0.3594
Q_8	0.2065	0.2766	0.2311	0.1975	0.2065	0.3642	0.3544	0.3416
Q_9	0.1742	0.2614	0.2200	0.1755	0.1742	0.3541	0.3432	0.3398

Source: Secondary and Simulation Data Processed (2022)