JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 6, No. 4, October 2022, pp. 1013-1026

Principal Component Regression Modelling with Variational Bayesian Approach to Overcome Multicollinearity at Various Levels of Missing Data Proportion

Nabila Azarin Balqis[1], Suci Astutik[2], Solimun[3]
[1,2,3]Departement of Statistics, University of Brawijaya, Indonesia
nabilaazarin24@gmail.com[1], suci_sp@ub.ac.id[2], solimun@ub.ac.id[3]

Article History:
Received : 30-07-2022
Revised : 17-09-2022
Accepted : 24-09-2022
Online : 08-10-2022

Keywords:
Missing Data;
Multi-collinearity;
Principal Component Analysis;
Principal Component Regression;
Variational Bayesian PCA.

ABSTRACT

This study aims to model Principal Component Regression (PCR) using Variational Bayesian Principal Component Analysis (VBPCA) with Ordinary Least Square (OLS) as a method of estimating regression parameters to overcome multicollinearity at various levels of the proportion of missing data. The data used in this study are secondary data and simulation data contaminated with collinearity in the predictor variables with various missing data proportions of 1%, 5%, and 10%. The secondary data used is the Human Depth Index in Java in 2021, complete data without missing values. The results indicate that the multicollinearity in secondary and original data can be optimally overcome as indicated by the smaller standard error value of the regression parameter for the PCR using VBPCA method which is smaller and has a relative efficiency value of less than 1. VBPCA can handle the proportion of missing data to less than 10%. The proportion of missing data causes information from the original variable to decrease, as evidenced by immense MAPE value and the parameter estimation bias that gets bigger. Then the cross validation (Q^2) value and the coefficient of determination ($adjusted\ R^2$) are get smaller as the proportion of missing data increases.

https://doi.org/10.31764/jtam.v6i4.10223

A. INTRODUCTION

One of the problems that often occurs in regression data is that there is a strong correlation between predictor variables so that the classical assumption or linear regression, namely non-multicollinearity, is prone to being violated. A strong correlation between predictor variables will cause parameter estimates using the least squares method to be obtained with a large effect of parameter estimator variance or even cannot be obtained (H. Kim & Jung, 2020). This problem can be solved by using Principal Component Regression (PCR) method which is a combination analysis between multiple linear regression and Principal Component Analysis (PCA). Suggestions regarding alternatives to the Ordinary Least Squares (OLS) estimator when the assumptions of mutually independent variables are not met are given to the PCR method (Ayinde et al., 2020).

In the process, PCA has a weakness, namely when faced with missing data. Missing data are observations that are not stored for a variable in the desired observation (Kang, 2013). Most researchers assume that missing observations do not intrinsically interfere with statistical

analysis of data sets (Marcelino et al., 2022), but it becomes a more critical problem when missing observations involve a multi-item instrument, due to the lack of information even in one of the data sets, the item which leads to the inability to calculate the total score of the instrument (Tsiampalis & Panagiotakos, 2020). Statistical analysis tends to be biased when more than 10% of data are missing (Bennet, 2001). 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) (Little & Rubin, 1987). 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 (Alruhaymi & Kim, 2021).

Some statistical analysis software cannot perform the PCA process if there is missing data, but some can also perform the process by removing sample rows on data that has missing observations. Some of the problems caused by missing data on statistical methods are causing bias in parameter estimators (Groenwold & Dekkers, 2020), reducing the strength of test statistics, and reducing sample representation so that the conclusions obtained are invalid. In regression data containing multicollinearity and missing data on the predictor variables, the classical PCR model can only overcome multicollinearity but not with missing data. Thus, its ability to deal with missing data is incomprehensible and becomes an important open challenge (Agarwal et al., 2021).

There are methods that can be used to handle multicollinearity and missing data, namely Probabilistic PCA (PPCA) and Variational Bayesian PCA (VBPCA). VBPCA was originally introduced with the aim of selecting the number of principal components which was then used to overcome missing data in PCA (Bishop, 1999). VBPCA is superior to PPCA because VBPCA can overcome the weakness of PPCA, namely the problem of overfitting. Handling overfitting problem can be handled by the VBPCA method through a Bayesian framework using a variational expectation-maximization (EM) iterative algorithm to search for each main sub-space which can then automatically select the optimal number of principal components (Li et al., 2020).

Several studies applying VBPCA to missing data have been conducted. Research by Yordani et al. (2015), which performs PCA on missing data using the VBPCA method. The data used in the research is simulation data and it is concluded that the VBPCA method can provide a high tolerance for missing data. In addition, research by Li et al. (2020) conducted a study on extracting common mode errors from regional GNSS (Global Navigation Satellite System) based on time series in the presence of missing data using VBPCA. The study concluded that the VBPCA method is a more efficient alternative method for extracting CME or general errors from the time series of regional GNSS positions in the presence of missing data. However, research on PCR using VBPCA as a PCA method to overcome multicollinearity and determine missing data limits has not been found.

This study aims to perform PCR modelling with the Variational Bayesian approach or VBPCA to overcome multicollinearity and missing data in the regression data. The parameter estimation method used is OLS which is the novelty in this study and has not been carried out in previous study. 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.

B. METHODS

The method used in this study is a combination of VBPCA with multiple linear regression analysis. Using OLS, the PCA method formed from The VBPCA method will be regressed with the response variable. The VBPCA method will discuss the multicollinearity and proportion of missing data that has been simulated. The data used in this study is secondary data and simulation data generated from secondary data. The secondary data used is Human Development Index from three provinces in Indonesia, namely East Java, West Java, and Central Java in 2021, totaling 100 observations.

Both classical PCR and PCR with the VBPCA approach will be carried out on secondary and simulation data. This study's secondary data are complete data without missing values. In contrast, the simulation data is data that contains multicollinearity and missing values on the predictor variables. There are nine predictor variables ($p = 9$) and one response variable in the secondary data which are used to generate the simulation data. Simulationdata which will be simulated with missing data with the percentage of missing data being 1%, 5%, and 10% of the total complete data.

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 (McDonald & Galarneau, 1975):

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

where $i = 1,2,\ldots,n$ and $j = 1,2,\ldots,k$. 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. The hypothesis for Little's MCAR test is given as follow (Little, 1988):

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

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

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

   $$d^2 = \sum_{j=1}^{J} n_j \left(\bar{\boldsymbol{y}}_{\boldsymbol{o}_j} - \hat{\boldsymbol{\mu}}_{\boldsymbol{o}_j}\right)^T \sum_{o_j}^{-1} \left(\bar{\boldsymbol{y}}_{\boldsymbol{o}_j} - \hat{\boldsymbol{\mu}}_{\boldsymbol{o}_j}\right) \tag{2}$$

   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.
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 e. 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 relative efficiency, bias, cross validation (Q^2), MAPE, and adjusted R^2.
9. Conclusion and suggestion.

C. RESULT 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 over-fitting which is the weakness of several combined methods when faced with different data sets.

1. Missing Data Mechanism Test and Non-Multicollinearity Test

Testing the assumption of missing data mechanism used Little MCAR test. The assumption test is carried out for all simulation data with α = 5%. The test results are presented as shown in Table 1.

Table 1. Little's MCAR Test Results

Proportion of Missing Data	Little's MCAR (d^2)	*p-value*
1%	90.4	0.071
5%	117	0.749
10%	137	0.692

Based on the test results in Table 1, 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. This means that the pattern of missing data that is formed occurs randomly and not systematically. The non-multicollinearity assumption test aims to see whether there is multicollinearity in the predictor variables of secondary data. VIF values for secondary data and simulation data are presented as shown in Table 2.

Table 2. VIF Value Results

Variable	Secondary Data	Simulation Data		
		Missing Data 1%	Missing Data 5%	Missing Data 10%
X_1	25.57	4,001.24	5,612.04	17,924.03
X_2	75.43	169.87	183.47	283.64
X_3	66.61	60.28	62.49	256.94
X_4	432.21	3,986.02	5,220.55	16,9274.70
X_5	1.89	2.76	3.92	6.39
X_6	25.76	3,058.11	3,647.52	8,258.32
X_7	2.01	3.01	4.98	13.49
X_8	3.87	2,753.78	3,428.42	7,239.02
X_9	1.99	146.24	154.23	365.12

Based on the VIF value in Table 2, it can be seen that the secondary data and all simulation data have more than one predictor variable that has a VIF value of more than 10. This shows that in the secondary data, the non-multicollinearity assumption is not met. In addition, the entire proportion of missing data in the simulation data has been contaminated with collinearity. So that the multicollinearity simulation on the predictor variables of the simulation data is carried out according to the expected scenario.

2. Estimation of Principal Component Regression Parameters

In this study, the focus of research is PCR modelling with the VBPCA method approach. However, the classical PCR method was also included as a comparison for the developed method. The difference between the two lies in the analysis step. In classical PCR method, it consists of two main steps, namely conducting PCA and regressing the principal components formed on the response variable. Whereas in PCR with the VBPCA method approach, the first step is to estimate the missing observations using a Variational Bayesian approach. After obtaining the complete data set, then the PCA method is carried out and regresses the principal components formed to the response variable.

In classical PCR, in conducting PCA, not all software in statistics can run the analysis process if there is missing data. In some other software, PCA method on data that has missing observations can be run. This is because the software provides special treatment if there is missing data, namely eliminating rows in the sample containing missing observations. So, by using the software, the missing data simulation of 1%, 5%, and 10% will lose the sample rows of 9, 41, and 83 samples. Therefore, the covariance matrix of the predictor variable cluster with the remaining sample for classical PCR method is obtained.

In PCR with the VBPCA method approach, the principal component in the VBPCA method uses a prior distribution which is used to estimating missing observations. 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, the missing observations can be estimated to obtain a complete data set.

After obtaining the complete data set, then the covariance matrix is obtained which is used to form the eigenvalues. The eigenvalues formed will be used to form eigenvectors called loading values. The value of loadings will then be formed by the principal component model. The main component model that is formed is then regressed to the response variable. The PCR model that was formed previously is a model that has a principal component coefficient, so it is substituted back into the PCR model and returned to the initial observation unit as in the secondary data. The results of the estimation of the regression parameters for the classical PCR method are given as shown in Table 3.

Table 3. Parameters Estimate of Classical PCR Method

Parameter	Secondary Data		Missing Data 1%		Missing Data 1%		Missing Data 1%	
	$\widehat{\boldsymbol{\beta}}$	$\left(\boldsymbol{S_e}(\widehat{\boldsymbol{\beta}})\right)$	$\widehat{\boldsymbol{\beta}}$	$\left(\boldsymbol{S_e}(\widehat{\boldsymbol{\beta}})\right)$	$\widehat{\boldsymbol{\beta}}$	$\left(\boldsymbol{S_e}(\widehat{\boldsymbol{\beta}})\right)$	$\widehat{\boldsymbol{\beta}}$	$\left(\boldsymbol{S_e}(\widehat{\boldsymbol{\beta}})\right)$
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

Based on the estimation results of the classical PCR method parameters formed in Table 3, the classical PCR model for secondary data and all simulation data is given in the following model.

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

$$\hat{Y} = 1{,}1175 - 2{,}3236X_1 - 1{,}1993X_2 - 0{,}0056X_3 + 2{,}2375X_4 - 8{,}6233X_6 + 7{,}9150X_8 + 0{,}4790X_9$$

The results of the estimation of the regression parameters for the PCR using VBPCA method approach are given as shown in Table 4.

Table 4. Parameters Estimate of PCR using VBPCA Approach

Parameter	Secondary Data		Missing Data 1%		Missing Data 1%		Missing Data 1%	
	$\widehat{\boldsymbol{\beta}}$	$\left(\boldsymbol{S_e}(\widehat{\boldsymbol{\beta}})\right)$	$\widehat{\boldsymbol{\beta}}$	$\left(\boldsymbol{S_e}(\widehat{\boldsymbol{\beta}})\right)$	$\widehat{\boldsymbol{\beta}}$	$\left(\boldsymbol{S_e}(\widehat{\boldsymbol{\beta}})\right)$	$\widehat{\boldsymbol{\beta}}$	$\left(\boldsymbol{S_e}(\widehat{\boldsymbol{\beta}})\right)$
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

Based on the estimation results of the PCR using VBPCA method approach parameters formed in Table 4, the classical PCR model for secondary data and all simulation data is given in the following model.

PCR using VBPCA Approach Model of 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 using VBPCA Approach Model of 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 using VBPCA Approach Model of 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 using VBPCA Approach Model of Simulation Data (Missing Data 10%)

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

3. Model Evaluation

The evaluation of the model in this study used six measures. The evaluation model used to see the effectiveness of the method in overcoming multicollinearity is the standard error of the regression parameter estimator and relative efficiency, then to see the effectiveness of the method in overcoming missing data, we used cross validation (Q^2), MAPE, and bias. Meanwhile, to evaluate the model to see the overall goodness of the model and to see how the PCR using VBPCA approach can overcome overfitting, we used the adjusted R^2 value.

a. Relative Efficiency

One measure of the goodness of the regression estimator can be seen through the relative efficiency. Relative efficiency is used to see the most efficient regression estimator in dealing with multicollinearity. The relative efficiency is calculated based on the variance of the regression parameter estimator. The variance of the regression parameter estimator is the square of the standard error of the regression parameter estimation. The standard error of 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 large so that the standard error of the regression parameter coefficient is also large. The standard errors of estimating the regression parameters are presented in Table 3 and Table 4.

Based on the standard error value of the regression parameter estimator in the secondary data, it can be seen that the results of the estimation of the standard error value of the regression parameter for the classical PCR and the PCR using VBPCA approach have the same value, because the secondary data has complete observations so that the missing observation estimation process is VBPCA method was not used. In the simulation data, the standard error value of the regression parameter estimator in the PCR using VBPCA approach is smaller than the classical PCR method. The PCR using VBPCA method approach

overcomes missing data, so that complete observations will be obtained that are in accordance with the information in the secondary data. In addition, 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 greater which causes the results of the analysis to be less valid. Furthermore, from the standard error of the regression parameter estimator, the relative efficiency can be formed. The results of the calculation of the relative efficiency are presented in Table 5. In the relative efficiency comparison table, three different colours are used, and each colour is divided into dark, normal, and light. The dark colour indicates that the relative efficiency is more than 1, while the light colour indicates that the relative efficiency is less than 1. For normal colour, it indicates that the relative efficiency is equal to 1. This is done to make it easier to read the table. As shown in Table 5.

Table 5. Comparison of the Relative Efficiency Value

$\boldsymbol{\beta_j}$	**Data**			
	Secondary Data	**Missing Data 1%**	**Missing Data 5%**	**Missing Data 10%**
β_0	1.00	0.87	0.51	0.29
β_1	1.00	187.92	44.71	0.30
β_2	1.00	0.96	0.17	0.33
β_3	1.00	107.77	11.72	0.04
β_4	1.00	2.28	0.49	0.56
β_5	1.00	1.39	0.53	0.81
β_6	1.00	2.44	0.19	0.00
β_7	1.00	3.52	0.79	0.13
β_8	1.00	1.52	0.46	0.03
β_9	1.00	0.45	0.11	0.24

Based on the calculation of the relative efficiency of $\hat{\beta}_{classical\ PCR}$ to $\hat{\beta}_{PCR-VBPCA}$ in Table 5, on the simulation data, it can be seen that the relative efficiency value of the classical PCR parameter estimator against the PCR using VBPCA method approach, the parameter estimator that has a relative efficiency value of more than 1 is less than the parameter estimator that has relative efficiency value less than 1. This shows that the parameter estimator of the PCR using VBPCA method approach is more efficient when compared to the classical PCR parameter estimator.

In secondary data, the relative efficiency of the classical PCR parameter estimator to the PCR using VBPCA approach has the same value, namely 1 for all parameters in the secondary data. This is because the classical PCR and PCR using VBPCA approach produce the same parameter estimates and models on secondary data that have complete observations. In addition, the higher the proportion of missing data, the fewer parameters that have a relative efficiency value of more than 1. This shows that the higher the proportion of missing data will reduce the efficiency of parameter estimation of a method. The amount of missing data will make the method less able to describe the data optimally.

b. Bias

One of the consequences of missing data is that it causes a bias in the parameter estimator. To see the effect of missing data on the model, we used the calculation of the average bias of the regression parameter estimator in this study. The results of the calculation of the average bias of the regression parameter estimators are presented as shown in Table 6.

Table 6. Average of Regression Parameter Estimator Bias

Simulation Data	Method	
	Classical PCR	**PCR using VBPCA Approach**
Missing Data 1%	0.1889	0.1328
Missing Data 5%	1.0330	0.4939
Missing Data 10%	2.8868	0.7277

Based on the results of the average bias of the regression parameters estimator in Table 6, it can be seen that the average bias of the parameter estimators of the PCR with VBPCA approach is smaller than the average bias of the classical PCR parameters estimators. This shows that the PCR using VBPCA approach can handle missing data better than the classical PCR method because missing data has been estimated. In addition, the higher the proportion of missing data in the data, the larger the regression parameter estimation bias will be. This shows that the amount of missing data in data has a big effect on the process of estimating the regression parameters, that is, it causes a larger bias so that the regression parameter estimation is not BLUE anymore.

c. Cross Validation (Q^2)

The optimal principal component in the classical PCR method as well as the VBPCA method will be indicated by the value of Q^2. The value of Q^2 from the analysis is presented in Table 7.

Table 7. Comparison of Cross Validation (Q^2)

Q (PC)	Method							
	Classical PCR				**PCR - 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

The comparison of the cross validation of the principal components in Table 7 is visualized in graphical form in Figure 1.

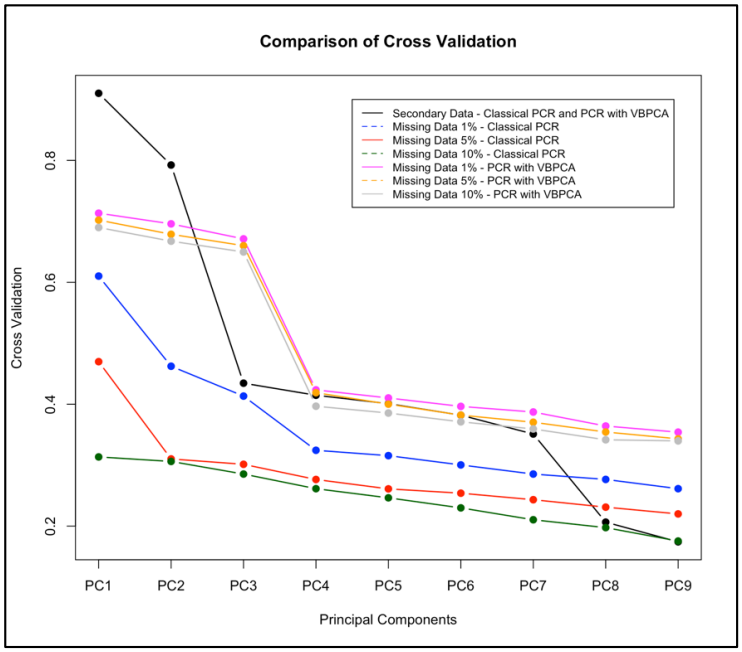

Figure 1. Cross Validation (Q^2) of Principal Components

Based on the cross validation values in Table 7 and Figure 2, it can be seen that in secondary data, the classical PCR method and PCR using VBPCA approach have the same Q^2 value caused by a complete set of observations. In addition, the secondary data has a Q^2 value which tends to be higher than other methods because there are no missing values, so that all the information formed by the principal components can explain the data optimally. Meanwhile, in the simulation data, the PCR using VBPCA approach has higher Q^2 value than the classical PCR method 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 can better explain the variance of 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.

d. Mean Absolute Percentage Error (MAPE)

In this study, the MAPE value was used to see the estimation results of missing data by the VBPCA method. The VBPCA method will be said to have good performance in overcoming missing observations if the resulting MAPE value is less than 10%. The smaller the resulting MAPE value, the better the VBPCA method in overcoming the missing observation values. The results of the MAPE calculation for the VBPCA method on the entire proportion of missing data are given in Table 8.

Table 8. Comparison of MAPE Value

Proportion of Missing Data	MAPE Value
1%	1.67%
5%	6.24%
10%	10.21%

Based on the MAPE calculation results for all 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%. While for missing data 10%, the MAPE value of the VBPCA method is 10.21% which is worth more than 10%. This shows that the VBPCA method can handle missing data until less than 10% of observations are missing from all complete data.

e. Adjusted R^2

Evaluation of the model to see the goodness of the PCR model formed in this study used adjusted R^2 value. 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 coefficient of determination for all methods are presented in Table 9.

Table 9. Comparison of Adjusted R^2

	Method	
Data	**Classical PCR**	**PCR - VBPCA**
Secondary Data	0.4912	0.4912
Missing Data 1%	0.3901	0.4057
Missing Data 5%	0.3867	0.4042
Missing Data 10%	0.3711	0.4024

The comparison of adjusted R^2 values in Table 9 is visualized in graphic form as shown in Figure 2.

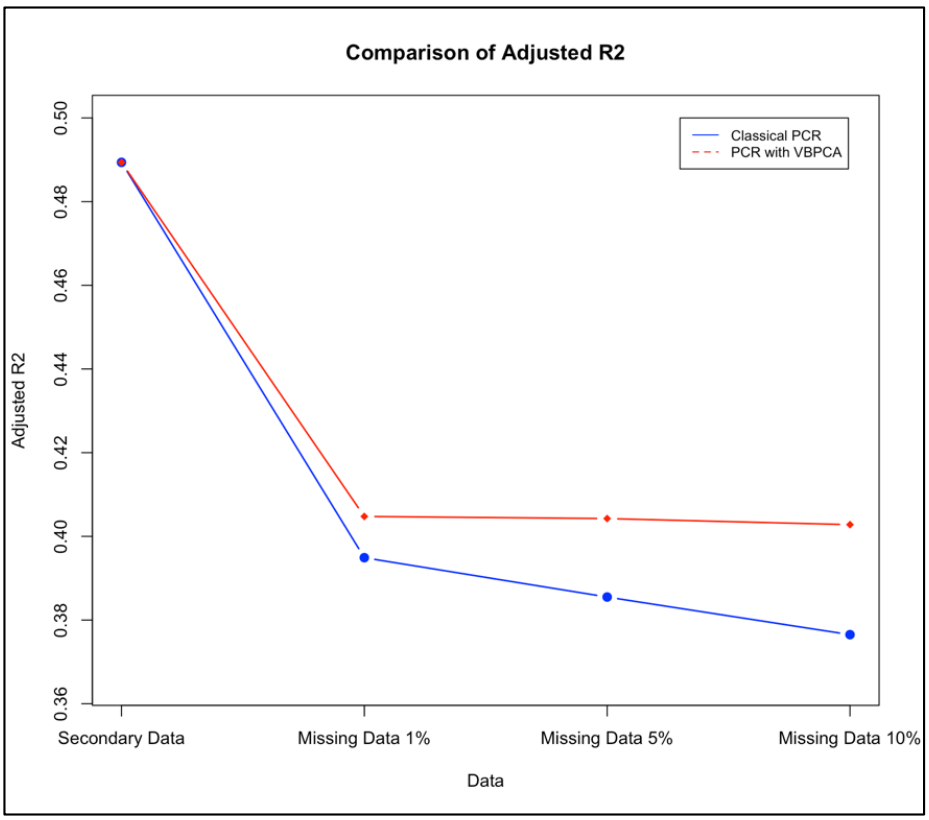

Figure 2. Comparison of Adjusted R^2

Based on the calculation results of the adjusted R^2 value in Table 9 and Figure 2, 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.

D. CONCLUSION AND SUGGESTIONS

Based on the results of research conducted in applied studies and simulation studies in this study, the following conclusions can be drawn: (1) The results of the estimation of the PCR parameters with VBPCA approach using OLS method have a constant value (the sign of the parameter coefficient) which is in accordance with the a priori theory on the proportion of missing data up to 5%. The results of the analysis show that the greater the proportion of missing data, the more unstable the estimation results of the regression parameters and the larger the standard error of the regression parameters; and (2) Based on the evaluation of the classical PCR and PCR with VBPCA approach, it can be concluded that: (a) the simulation data with 10% missing data proportion has a MAPE value of 10.21% which is greater than 10%. This exceeds the MAPE value limit, so the VBPCA method can handle missing data of less than 10%; (b) the proportion of missing data has an effect on the size of the model's goodness. The higher the proportion of missing data, the less optimal the model will be. This can be seen through the larger MAPE value and parameter estimator bias, then seen through the smaller cross validation (Q^2) and adjusted R^2 values. The proportion of missing data results in the information generated from the predictor variables being less than optimal; and (3) the PCR model with VBPCA method approach is able to overcome overfitting when faced with different data sets. This is evidenced by the adjusted R^2 value which is the same as the classical PCR method on secondary data that has complete data.

Based on the results in this study, a large sample was used in this study, so that in future research it can be recommended to use a small sample so that it can determine the effectiveness of the PCR with VBPCA method approach in overcoming multicollinearity and missing data in small samples. In addition, this study resulted in a small adjusted R^2 value, so that in future research it can be considered again the selection of predictor variables used and other

combinations of methods can be used to overcome the multicollinearity and missing data in the regression data which may increase the adjusted R^2 value. The results of this study differ from the research conducted by Yordani's (Yordani et al., 2015), where if the VBPCA method is used alone without a combination of other methods, the VBPCA can handle missing data up to 35%. Whereas in this study, if VBPCA is combined with the regression method, VBPCA can only handle missing data of less than 10%.

ACKNOWLEDGEMENT

We are very grateful to experts for their appropriate and constructive suggestions to improve this template. The author expresses appreciation and thanks to Mrs. Suci Astutik as supervisor 1 and Mr. Solimun as supervisor II who have helped in writing this paper. The author would also like to thank her family who have always supported him in completing this paper.

REFERENCES

Agarwal, A., Shah, D., Shen, D., & Song, D. (2021). On Robustness of Principal Component Regression. *Journal of the American Statistical Association*, *116*(536), 1731–1745. https://doi.org/10.1080/01621459.2021.1928513

Ahmad, A. U., Balakrishnan, U. V., & Jha, S. (2021). A Study of Multicollinearity Detection and Rectification under Missing Values. *Turkish Journal of Computer and Mathematics Education,* 12(1), 399-418. https://doi.org/10.17762/turcomat.v12i1s.1880

Alabi, O. O., Ayinde, K., Babalola, O. E., Bello, H. A., & Okon, E. C. (2020). Effects of Multicollinearity on Type I Error of Some Methods of Detecting Heteroscedasticity in Linear Regression Model. *Open Journal of Statistics*, *10*(04), 664–677. https://doi.org/10.4236/ojs.2020.104041

Alruhaymi, A. Z., & Kim, C. J. (2021). Study on the Missing Data Mechanisms and Imputation Methods. *Open Journal of Statistics*, *11*(04), 477–492. https://doi.org/10.4236/ojs.2021.114030

Arumsari, M., Tri, A., & Dani, R. (2021). Peramalan Data Runtun Waktu Menggunakan Model Hybrid Time Series Regression-Autoregressive Integrated Moving Average. In *Jurnal Siger Matematika* (Vol. 02, Issue 01). http://dx.doi.org/10.23960%2Fjsm.v2i1.2736.

Ayinde, K., Lukman, A. F., Alabi, O. O., & Bello, H. A. (2020). A New Approach of Principal Component Regression Estimator with Applications to Collinear Data. *International Journal of Engineering Research and Technology*, *13*(7), 1616–1622. https://doi.org/10.37624/ijert/13.7.2020.1616-1622

Bennet, D. A. (2001). How Can I Deal with Missing Data in My Study? *Aust N Z J Public Health, 25*(5), 464-469. DOI: 10.1111/j.1467-842X.2001.tb00294.x

Bishop, C. M. (1999). Variational Principal Components. *Ninth International Conference on Artificial Neural Networks, ICANN, IEE, Vol. 1,* 509-514. https://doi.org/10.1049/CP:19991160.

Schipper, N. C., & Deun, K. V. (2021). Model Selection Techniques for Sparse Weight-Based Principal Component Analysis. *Journal of Chemometrics*, *35*(2). https://doi.org/10.1002/cem.3289

Diah, S., Larasati, A., Nisa, K., Setiawan, D. E., Soemantri Brojonegoro, J., & Lampung, B. (2020). Analisis Regresi Komponen Utama Robust dengan Metode Minimum Covariance Determinant-Least Trimmed Square (MCD-LTS). *Jurnal Siger Matematika*, 1(1), 1-9. http://dx.doi.org/10.23960%2Fjsm.v1i1.2472

Estrada, Ma. del R. C., Camarillo, M. E. G., Parraguirre, M. E. S., Castillo, M. E. G., Juárez, E. M., & Gómez, M. J. C. (2020). Evaluation of Several Error Measures Applied to the Sales Forecast System of Chemicals Supply Enterprises. *International Journal of Business Administration*, *11*(4), 39. https://doi.org/10.5430/ijba.v11n4p39

Groenwold, R. H. H., & Dekkers, O. M. (2020). Missing Data: The Impact of What is Not There. *European Journal of Endocrinology*, *183*(4), E7–E9. https://doi.org/10.1530/EJE-20-0732

Jollife, I. T., & Cadima, J. (2016). Principal Component Analysis: A Review and Recent Developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (Vol. 374, Issue 2065). https://doi.org/10.1098/rsta.2015.0202

Kang, H. (2013). The Prevention and Handling of the Missing Data. *Korean Journal of Anesthesiology* (Vol. 64, Issue 5, pp. 402–406). https://doi.org/10.4097/kjae.2013.64.5.402

Karch, J. (2020). Improving on Adjusted R-squared. *Collabra: Psychology*, *6*(1). https://doi.org/10.1525/collabra.343

Kim, H., & Jung, H. Y. (2020). Ridge Fuzzy Regression Modelling for Solving Multicollinearity. *Mathematics*, *8*(9). https://doi.org/10.3390/math8091572

Kim, S., & Kim, H. (2016). A New Metric of Absolute Percentage Error for Intermittent Demand Forecasts. *International Journal of Forecasting*, *32*(3), 669–679. https://doi.org/10.1016/j.ijforecast.2015.12.003

Li, W., Jiang, W., Li, Z., Chen, H., Chen, Q., Wang, J., & Zhu, G. (2020). Extracting Common Mode Errors of Regional GNSS Position Time Series in the Presence of Missing Data by Variational Bayesian Principal Component Analysis. *Sensors (Switzerland)*, *20*(8). https://doi.org/10.3390/s20082298

Liantoni, F., & Agusti, A. (2020). Forecasting Bitcoin Using Double Exponential Smoothing Method Based on Mean Absolute Percentage Error. *International Journal on Informatics Visualization, 4*(2). https://doi.org/10.20630/joiv.4.2.335

Little, R. J. A. & Rubin, D. B. (1987). Statistical Analysis with Missing Data. Hoboken: John Wiley and Sons.

Little, R. J. A. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*, *83*(404), 1198–1202. https://doi.org/10.1080/01621459.1988.10478722

Mahmoudi, M. R., Heydari, M. H., Qasem, S. N., Mosavi, A., & Band, S. S. (2021). Principal Component Analysis to Study the Relations Between the Spread Rates of COVID-19 in High Risks Countries. *Alexandria Engineering Journal*, *60*(1), 457–464. https://doi.org/10.1016/j.aej.2020.09.013

Marcelino, C. G., Leite, G. M. C., Celes, P., & Pedreira, C. E. (2022). Missing Data Analysis in Regression. *Applied Artificial Intelligence*. https://doi.org/10.1080/08839514.2022.2032925

McDonald, G. C., & Galarneau, D. I. (1975). A Monte Carlo Evaluation of Some Ridge-type Estimators. *Journal of the American Statistical Association*, *70*(350), 407–416. https://doi.org/10.1080/01621459.1975.10479882

Astivia, O. L. O. & Zumbo, B. D. (2019). 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*, *24*. https://doi.org/10.7275/q5xr-fr95

Pham, H. (2019). A New Criterion for Model Selection. *Mathematics*, *7*(12), 1215. https://doi.org/10.3390/MATH7121215

Rutledge, D. N., Roger, J.-M., & Lesnoff, M. (2021). Different Methods for Determining the Dimensionality of Multivariate Models. *Frontiers in Analytical Science*, *1*. https://doi.org/10.3389/frans.2021.754447

Tsiampalis, T., & Panagiotakos, D. B. (2020). 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*, *20*(1). https://doi.org/10.1186/s12874-020-01038-3

Wulandari, S., Salam, N., & Anggraini, D. (2010). Perbandingan Metode Robust MCD-LMS, MCD-LTS, MVE-LMS, dan MVE-LTS dalam Analisis Regresi Komponen Utama. *Jurnal Matematika Murni dan Terapan, 4*(1), 57-64.

Yordani, R. (2015). Penerapan Model Inferemsi Bayesian dengan Variational Bayesian Principal Component Analysis (VBPCA) dalam Mengatasi Missing Data Analisis Komponen Utama*. Jurnal Aplikasi Statistika & Komputasi Statistik, 7*(2), 51-69. https://doi.org/10.34123/jurnalasks.vbi1.12

Ziegel, E. R. (1991). Linear Statistical Models: An Applied Approach. *Technometrics*, *33*(2), 248–248. https://doi.org/10.1080/00401706.1991.10484830