International Journal of Science and Mathematics Education
Vol. 1, No.3 September 2024
e-ISSN: 3062-7214; p-ISSN: 3062-7206, Hal 13-23
DOI ; https://doi.org/10.62951/ijsme.v1i3.25
Available online at: https://international.arimsi.or.id/index.php/IJSME

Elastic Net Principal Component Regression With an Application

Afraa A. Hamada
Department of Statistics- College of Administration and Economics,
University of Al- Qadisiyah, Iraq

Author correspondence: afraa.hamada@qu.edu.iq

Abstract. *To overcome the difficulties of high-dimensional data, Elastic Net Principal Component Regression (ENPCR), a potent statistical technique, combines Elastic Net regularization with Principal Component regression (PCR). When dealing with Multicollinearity among predictors, this method is especially helpful because it enables efficient variable selection while preserving interpretability. PCA is initially used in ENPCR to reduce the dataset's dimensionality by converting correlated variables into a group of uncorrelated principal components. The Elastic Net regression model then uses these elements as inputs and penalizes the regression coefficients using both L1 and L2 penalties. By promoting sparsity, this dual regularization lessens overfitting and helps the model concentrate on its most important components. simulated studies and Real datasets are used to demonstrate the our proposed method .*

Keywords: *regularization method, Elastic Net technique, principal component regression, simulation scenario*

1. INTRODUCTION

Datasets classified as high-dimensional have a high number of variables relative to observations(Fan, J., & Lv, J. (2008)). High-dimensional data becomes complex in terms of analysis and interpretation.to overcame this problem maybe through using one regularization method(Bühlmann, P., & van de Geer, S. (2011)). It is known that as the dimensions of the data in regression analysis increase, the statistical problems arise, providing us with inefficient estimators. One of the most important of these problems is the issue of Multicollinearity((Montgomery, D *et al ,2012)*). In regression analysis, a situation known as Multicollinearity occurs when there is a significant correlation between two or more independent variables, producing redundant results. Changes in one predictor variable are linked to changes in another when Multicollinearity is present. The variances of the coefficient estimations may be inflated by this high correlation, producing results that are unstable and unreliable (O'Brien, R. G. (2007)) . Regression analysis Multicollinearity can result in inaccurate coefficient estimates and difficult-to-understand model interpretation. To lessen the problems related to Multicollinearity, Elastic Net Principal Component Regression (ENPCR) is a useful strategy that combines the advantages of Elastic Net regularization and Principal Component regression (PCR) (Hastie, T et al2009). Our proposed method(ENPCR) is considered a distinctive approach for addressing all types of perfect and semi-perfect Multicollinearity. This is achieved through the advantages offered by our method, which combines two effective techniques for handling Multicollinearity. Additionally, it reduces high

Received: July 18, 2024; Revised: August 19, 2024; Accepted: September 21, 2024; Published: September 25, 2024

dimensions to a manageable number that is interpretable. The remainder of the paper is organized as follows: Principal Component Regression is presented in Section 2. In section 3,we introduced our proposed method Elastic Net principal component regression (ENPCR). In section 4, we tested the performance proposed method via simulation scenarios and medical real data. In section 5, we introduced a brief conclusions and recommendations.

2. LITERATURE REVIEW

Principal Component Regression

The main objective of regression models is to evaluate the causal relationship between a dependent variable and one or more predictive variables (Gelman, A., & Hill, J. (2007)). This model can be described by the following mathematical expression:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots + \beta_p x_p + \epsilon_i \qquad [i = 1,2 \dots p] \qquad (1)$$

$y_i$is the response variable, β_0 is intercept , $\beta_1, \beta_2 \dots \beta_p$ are the parameters of model linked with the predictive variables$(x_1, x_2, \dots\dots x_p)$.

ϵ_i is the random error term distributed according normal distribution with mean (0) and variance σ^2. We can introduced Multiple regression model by matrix approach as following :

$$Y = X\beta + \epsilon \qquad (2)$$

However, many regression models face certain statistical issues, making it very difficult to obtain efficient and reliable estimates for those models. One of the most significant issues is the problem of Multicollinearity, which directly affects the inflation of the variance of the estimated coefficients to a considerable extent (Montgomery, D. C.*et al2012).* . The Multicollinearity problem appears with regression models under several conditions (Hair, J. F., D. C.*et al 1998)*: Multicollinearity results from a high degree of correlation between two or more independent variables, making it challenging to discern the unique effects of each variable on the dependent variable. Perfect Multicollinearity results when one predictor variable may be described as a linear combination of other variables((O'Brien, R. G. (2007)). To cut down on Multicollinearity, choose and remove one of the linked predictors. In order to minimize the number of predictors while preserving important information, create composite variables by summing or averaging associated variables((Hair, J. F., D. C.*et al 1998)*). All the methods mentioned above are basic approaches that require a considerable amount of time to address the issue of multicollinearity. These methods may complicate the model and have adverse effects on the estimates. To overcome this problem, Principal Component Regression (PCR) can be used. It focuses on changing the initial explanatory variables, which have some degree of connection, into new variables, called principle components, that are linear

combinations of the initial variables without changing any of the original variables(Massy, W. F. (1965)). Comprehensive information regarding the observations of the original explanatory variables will be provided by these principal components. As indicated by the following equation, each main components linear combination may be formed by combining the explanatory variables ($x_1, x_2, \ldots . x_p$).

$$Z = XC \qquad (3)$$

Z is represented a matrix of principal components will be of dimensions($n * K$) .

X is represented a matrix of independent variable will be of dimensions($n * p$) .

C is represented a matrix of loadings or coefficients that specifies the contribution of the independent variable variables to the principal components. It will be of dimensions($p * k$) . It is orthogonal matrix of eigenvectors corresponding to the eigenvalues of the matrix($X^T X$). The C matrix is Diagonal matrix, ($CC^T = C^T C = I$)(Jolliffe, I. T. (2002)).

From use the feature of C matrix then equation (2) is become as following:

$$ZC^T = XCC^T \qquad (4)$$

where $CC^T = I_P$, Then X= ZC^T: In essence, it demonstrates how a linear combination of the principal components represented the independent variable((Lee, H.*et al 2015)*) . The model represented in the equation below illustrates the causal relationship between the dependent variable and a set of principal components.

$$Y = ZC^T\beta + \epsilon \qquad (5)$$

Where the β is a vector of parameters belong to the principal components.

we assumed the $ZC^T = \delta$, then the principal component regression take the following formula(Perez, L. V. (2017)):

$$Y = \delta\beta + \epsilon \qquad (6)$$

The estimates of the model represented in the equation(6) can be obtained by minimizing the following function:

$$\hat{\beta} = min \sum_{I=1}^{P} (y_i - \delta_i\beta)^2 \qquad (7)$$

Take the derivative of the sum of squared residuals with respect to β and set it to zero to obtain parameters estimation(Jolliffe, I. T. (2002)).

$$\hat{\beta} = (\delta^T\delta)^{-1}\delta Y \qquad (8)$$

The equation (8) They are intended for estimating the parameters of the principal component regression model, which is a good tool for addressing Multicollinearity. However,

when using regularization methods with principal component regression, we obtain a very robust method.

Elastic Net principal component regression

Elastic Net is a regularization strategy that improves variable selection and prediction performance in high-dimensional datasets by combining the advantages of both Lasso (L1 regularization) and Ridge (L2 regularization) regression methods. By adding a penalty to the sum of the squared coefficients, it overcomes the drawbacks of Lasso, especially its propensity to choose one variable from a set of highly correlated predictors. Two hyper parameters are introduced by the Elastic Net technique, one of which controls the L1 penalty and the other the L2 penalty. This allows for a flexible trade-off between coefficient shrinkage and variable selection. This method works particularly well when there are more predictors than observations or when there is a significant degree of correlation between the predictors. Elastic Net allows practitioners to increase prediction accuracy while obtaining more stable and interpretable models. When mixture the Elastic Net technique with the principal component regression we obtained the following :

$$\hat{\beta} = min \sum_{I=1}^{P} (y_i - \delta_i \beta)^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \qquad (9)$$

$\lambda_1 and\, \lambda_2$ are shrinkage parameter, where $0 \leq \lambda_1 and\, \lambda_2 \leq 1$. And the term $\sum_{I=1}^{P} (y_i - \delta_i \beta)^2$ is the check function of principal component regression, the term$\lambda_1 \|\beta\|_1$ By encouraging some coefficients to be exactly zero, the L1 penalty term encourages sparsity in the model. the term $\lambda_2 \|\beta\|_2^2$decreasing coefficients, helps to stabilise the estimates and reduces problems associated with Multicollinearity. The condition that connects the gradients of the residuals and the penalties is obtained by computing the gradient of the objective function and setting it to zero. To determine the ideal coefficients β, numerical methods are frequently utilized due to the difficulty posed by the L1 term. Combining Elastic Net regularization with principal component analysis's (PCA) advantages is Elastic Net Principal Component Regression, or ENPCR. Multicollinearity among predictors can make traditional regression methods difficult to utilize, which is why this methodology is especially helpful in high-dimensional data circumstances.in our proposed method ,we introduced a new methods used for treat Multicollinearity in regression model. Based on the functions (glmnet)(caret) in R programm, it is possible to build a specialized program for estimating the parameters and selecting variables in a principal component regression model by employing one of the regularization functions, we can building special R program of our proposed method.

Simulation Approach

Through simulation tests, the performance of our proposed method the (ENPCR)is examined. These investigations aim to assess the our proposed method robustness and efficacy in a range of scenarios. we will compared our proposed method with two of existing methods in same filed. First methods is(principle component regression) denoted by (P.C.R) which, it proposed by (Wehrens, R., & Mevik, B. H. (2007))within R package (pls) function (pcr) .Second method is (Bayesian lasso principal component regression with an application) denoted by(B.L.P.C.R) which it is proposed(AL-Sharoot, M. H., Kazem *et al 2023*),within special R package. In this simulation study, We will use four types of correlation coefficients between the independent variables as follows: $\rho_1 = 0.55,\ \rho_2 = 0.85\ , \rho_3 = 0.95\ and \rho_4 = 0.99$. Also,we used four sample size ($n_1 = 50, n_2 = 100, n_3 = 150\ and\ n_4 = 200,$) For comparison, two criteria were used: median of mean absolute deviations (MMAD), where $\text{MMAD} = median\left(mean\left(\left|\delta^T\hat{\beta} - \delta^T\beta^{true}\right|\right)\right)$and Standard deviation(S.D). our algorithm runs 11000 iterations , with the first 1,000 iterations eliminated for burn-in. In this simulation examples, two simulation scenarios are implemented as following:

First Simulation Scenario

The effectiveness of our proposed method (ENPCR) with sparse model, was demonstrated in this simulation scenario. Sparse model is take the following formula:

$$y_i = 1x_{1i} + 1x_{2i} + 2x_{5i} + 1x_{9i} + \varepsilon_i, \qquad \text{i=1,2,….200}$$

Nine independent variables was generated from a multivariate normal distribution $X \sim N_9(0, \Sigma_9)$ with mean zero (a vector of 0) and Covariance matrixΣ_p. $\beta = (0,1,1,0,0,2,0,0,0,0)$ is the actual coefficient of the explanatory variables, which includes the intercept term. The results of MMADs and standard deviations (SD) which listed in Table 1. It is readily noted that in all sample size and all correlation levels under considerations. Therefore, our proposed method(ENPCR) have a good performance compared with other two methods (P.C.R)(B.L.P.C.R). This is evident from the values of (MMAD)(S.D) calculated using our proposed method, which is much smaller than the values of (MMAD)(S.D) calculated using the comparison methods (P.C.R)(B.L.P.C.R).

Table 1. The MMADs and standard deviations for first Simulation					
Methods	**Methods**	$\rho_1 = 0.55$	$\rho_2 = 0.85$	$\rho_3 = 0.95$	$\rho_4 = 0.99$
n=50	P.C.R	0.871 (0.432)	0.767 (0.392)	0.881 (0.463)	0.743 (0.282)
	B.L.P.C.R	0.634 (0.303)	0.710 (0.383)	0.735 (0.376)	0.653 (0.330)
	ENPCR	**0.492 (0.218)**	**0.561 (0.333)**	**0.619 (0.349)**	**0.510 (0.268)**
n=100	P.C.R	0.963 (0.517)	0.869 (0.482)	0.934 (0.451)	0.887 (0.401)
	B.L.P.C.R	0.847 (0.491)	0.732 (0.369)	0.836 (0.428)	0.755 (0.384)
	ENPCR	**0.561 (0.267)**	**0.495 (0.152)**	**0.632 (0.362)**	**0.492 (0.123)**
n=150	P.C.R	1.124 (0.521)	1.294 (0.781)	1.452 (0.784)	1.307 (0.565)
	B.L.P.C.R	1.023 (0.511)	1.185 (0.652)	1.394 (0.463)	1.123 (0.562)
	ENPCR	**0.841 (0.419)**	**0.944 (0.454)**	**0.854 (0.434)**	**0.723 (0.350)**
n=200	P.C.R	1.318 (0.798)	1.429 (0.743)	1.521 (0.759)	1.451 (0.693)
	B.L.P.C.R	1.134 (0.674)	1.203 (0.564)	1.126 (0.657)	1.088 (0.519)
	ENPCR	0.864 (0.381)	0.741 (0.381)	0.863 (0.467)	0.736 (0.373)

Note: In the parentheses are Standard deviation(S.D).

The stability of the estimation algorithm is considered a very important aspect for assessing the superiority of the estimators according to the proposed method. In our ongoing research, we will rely on the plot of (**trace plot**) to illustrate the stability of the estimation algorithm as follows:

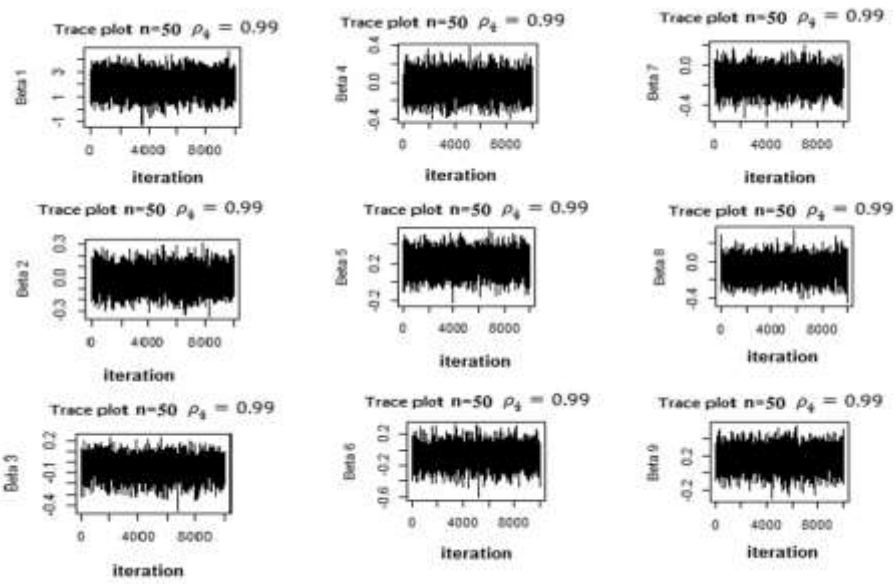

Figure (1) show the trace plot of ENPCR in sparse case and sample size is 50

$\boldsymbol{\rho_4 = 0.99}$

From the above figure, we see the our algorithm is very stationary via nine coefficients estimation. From this results, This result can be generalized to all estimators of the model (sparse) across all sample sizes and levels of correlation shown in the first simulation. Therefore, our algorithm is stable.

Second simulation scenario

The effectiveness of our proposed method (ENPCR) with dense model, was demonstrated in this simulation scenario. dense model is take the following formula:

$$y_i = 0.85x_{1i} + 0.85x_{2i} + 0.85x_{3i}+0.85x_{4i} + 0.85x_{5i} + 0.85x_{6i} + 0.85x_{7i}+0.85x_{8i} + 0.85x_{9i} + \varepsilon_{i_i}, \quad i=1,2,\ldots.200$$

Nine independent variables was generated from a multivariate normal distribution $X \sim N_9(0, \Sigma_9)$ with mean zero (a vector of 0) and Covariance matrixΣ_p. $\beta = (0,0.85,0.85,0.85,0.85,0.85,0.85,0.85,0.85,0.85)$ is the actual coefficient of the explanatory variables, which includes the intercept term. The results of MMADs and standard deviations (SD) which listed in Table 2. It is readily noted that in all sample size and all correlation levels under considerations. Therefore, our proposed method(ENPCR) have a good performance compared with other two methods (P.C.R)(B.L.P.C.R). This is evident from the values of (MMAD)(S.D) calculated using our proposed method, which is much smaller than the values of (MMAD)(S.D) calculated using the comparison methods (P.C.R)(B.L.P.C.R).

Table 2. The MMADs and standard deviations for first Simulation

Methods	**Methods**	$\rho_1 = 0.55$	$\rho_2 = 0.85$	$\rho_3 = 0.95$	$\rho_4 = 0.99$
n=50	P.C.R	1.239 (0.672)	1.383 (0.720)	1.186 (0.673)	1.341 (0.730)
	B.L.P.C.R	0.945 (0.407)	0.859 (0.311)	0.734 (0.286)	0.973 (0.363)
	ENPCR	**0.582 (0.197)**	**0.597 (0.210)**	**0.479 (0.238)**	**0.392 (0.107)**
n=100	P.C.R	1.188 (0.653)	1.049 (0.536)	1.022 (0.428)	0.961 (0.505)
	B.L.P.C.R	0.962 (0.392)	0.863 (0.357)	0.793 (0.369)	0.688 (0.463)
	ENPCR	**0.621 (0.358)**	**0.586 (0.422)**	**0.485 (0.174)**	**0.448 (0.281)**
n=150	P.C.R	1.275 (0.664)	1.117 (0.577)	1.106 (0.469)	1.232 (0.682)
	B.L.P.C.R	0.942 (0.517)	0.876 (0.439)	0.851 (0.392)	0.829 (0.380)
	ENPCR	**0.752 (0.363)**	**0.524 (0.266)**	**0.520 (0.305)**	**0.510 (0.275)**
n=200	P.C.R	1.117 (0.943)	1.043 (0.676)	1.006 (0.345)	0.922 (0.362)
	B.L.P.C.R	1.057 (0.792)	1.067 (0.452)	1.018 (0.383)	1.043 (0.561)
	ENPCR	0.624 (0.286)	0.549 (0.302)	0.493 (0.124)	0.457 (0.231)

Note: In the parentheses are Standard deviation(S.D).

The stability of the estimation algorithm is considered a very important aspect for assessing the superiority of the estimators according to the proposed method. In our ongoing research, we will rely on the plot of (**trace plot**) to illustrate the stability of the estimation algorithm as follows:

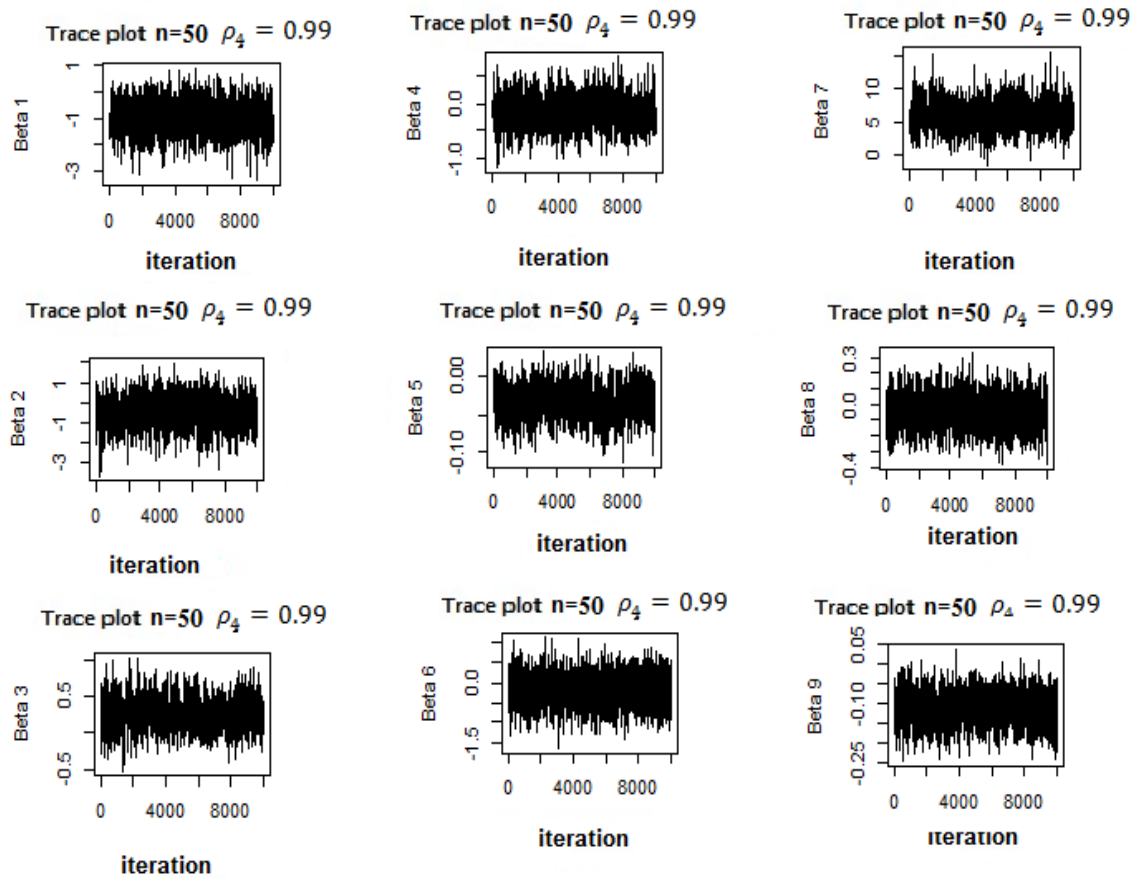

Figure (2) show the trace plot of ENPCR in dense case and sample size is 50 $\rho_4 = 0.99$

From the above figure, we see the our algorithm is very stationary via nine coefficients estimation. From this results, This result can be generalized to all estimators of the model (sparse) across all sample sizes and levels of correlation shown in the first simulation. Therefore, our algorithm is stable.

3. RESULTS AND DISCUSSION

Real Dataset

After testing and demonstrating the good performance of our method compared to the comparison methods using simulation techniques, Additionally, the behavior of our proposed method will be studied using medical data obtained from the maternity hospital in Diwaniya, with a sample size of 220 observations. Our study contain one response variable is represented Weight of a Newborn Baby (y_i),and nine independent variables are Gestation Duration (in days) (x_1), Number of Births(x_2), Is the mother diabetic? (x_3), Is the mother suffering from pregnancy-induced hypertension? (x_4), Mother's Age at Birth(x_5), Mother's Weight at Birth(x_6), Is the mother infected with COVID-19? (x_7), Is the mother a smoker? (x_8), Type of birth(x_9), After encoding the dependent variable and the independent variables, they were input into our proposed algorithm, where 11,000 iterations were executed, and the first 3,000 iterations were discarded to achieve more stable estimates .To compare our proposed method with previous approaches, we will employ two criteria: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), as well as the standard division

Table 3 show the values of RMSE, MAE and S.D for real data

Methods	RMSE	MAE	S.D
P.C.R	3.452	1.858	0.782
B.L.P.C.R	3.763	1.939	0.810
ENPCR	1.483	1.217	582

From the results presented in the table above, we notice that the (RMSE) calculated using our proposed method is much smaller than the (RMSE) calculated using the other two methods. Therefore, our proposed method demonstrates good performance in estimating parameters and selecting variables, even in the presence of Multicollinearity issues. We also observe the same result when using the criteria (MAE) and (S.D). After demonstrating that our proposed method has superiority, it will be used to analyze this medical data as follows:

Table 4 Show point estimation and interval estimation of parameters to real data

independent Variables	Name Variables	Point estimation	Interval estimation	
			lower	Upper
X_1	Gestation Duration (in days)	2.189	1.083	3.305
X_2	Number of Births	0.000	-0.002	0.002
X_3	Is the mother diabetic	0.016	0.000	0.023
X_4	Is the mother suffering from pregnancy-induced hypertension?	0.263	0.156	0.372
X_5	Mother's Age at Birth	2.281	1.413	2.843
X_6	Mother's Weight at Birth	0.000	-0.008	0.001
X_7	Is the mother infected with COVID-19?	0.035	-0.120	0.191
X_8	Is the mother a smoker?	0.000	-0.012	0.016
X_9	Type of birth	0.006	0.013	0.030

After estimating the coefficients of the model, we find that there are 3 independent variables that are not significant in building our model and can be excluded from its composition. However, on the other hand, there are 6 important variables that have different effects on the response variable (Weight of a Newborn Baby).

4. CONCLUSIONS AND RECOMMENDATION

Conclusions

We conclude that our proposed method demonstrates a good and superior performance in reducing high-dimensional data by combining two techniques used for this purpose. By mixing Principal Component Regression with Elastic Net, we achieve a robust and effective approach for handling high-dimensional datasets. Moreover, the proposed method is effective in addressing the issue of Multicollinearity, as well as the problem of an increasing number of independent variables relative to the sample size. Also, By concentrating on principal components that capture the greatest variance, the combination of PCA with Elastic Net improves the interpretability of the model and facilitates understanding of the relationships between the variables. Finally, Elastic Net Principal Component Regression (PCR) combines the strengths of Principal Component Analysis (PCA) and Elastic Net regularization, resulting in a robust model that effectively handles Multicollinearity and high-dimensional data.

Recommendation

We recommended by employing our prosed method (Elastic Net Principal Component Regression) is a potent and efficient method that strikes a compromise between simplicity, interpretability, and performance while analyzing high-dimensional datasets. When working with datasets where there are more predictors than observations, we recommended use Elastic Net PCR since it is a good way to control Multicollinearity and minimize overfitting. Expanding our current study to include other regularization methods with strong properties and combining them with Principal Component Regression can lead to robust models capable of addressing Multicollinearity and high-dimensional data simultaneously and efficiently.

REFERENCE

AL-Sharoot, M. H., Kazem, W. A. A. H., & Al-Fatlawi, S. A. (2023). Bayesian lasso principal component regression with an application. *Journal of Al-Qadisiyah for Computer Science and Mathematics*, *15*(4), 32-38.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer.

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(5), 849-911.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed.). Prentice Hall.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Jolliffe, I. T. (2002). *Principal component regression* (2nd ed.). Springer.

Lee, H., Park, Y. M., & Lee, S. (2015). Principal component regression by principal component selection. *Communications for Statistical Applications and Methods*, *22*(2), 173-180.

Massy, W. F. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, *60*(309), 234-256.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (5th ed.). Wiley.

O'Brien, R. G. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality and Quantity*, *41*(5), 673-690.

Perez, L. V. (2017). Principal component analysis to address multicollinearity. Whitman College: Walla Walla, WA, USA, 99362.

Wehrens, R., & Mevik, B. H. (2007). The pls package: Principal component and partial least squares regression in R.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301-320.