Supervised Principal Component Regression Analysis Of Multicollinear Data

By

Okonkwo, Evelyn Nkiruka

Nnamdi Azikiwe University, Awka, Nigeria

Okeke, Joseph Uchenna

Anambra State University, Uli, Nigeria

Nwabueze, Joy Chioma.

Micheal Okpala University of Agriculture, Umudike Nigeria

Abstract

This study has examined the performance of SPCR when there is presence of multicollinearity in the data. The situations where the predictor variables are correlated and when the number of predictors exceeds the number of observations were considered using both simulated and real data sets. The adequacy of SPCR model was compared with the classical principal component regression, and stepwise regression using AIC and SIC. The results obtained show that stepwise regression performed poorly in both simulated and real data sets. SPCR and classical principal component regression compete favourably to stepwise regression.

1.0 Introduction

It is a well known fact that ordinary least squares regression coefficients estimators may perform poorly when there is multicollinearity in the X the matrix of predictor variables and where the number of predictors greatly exceeds the number of observations. The variance of the components of the ordinary least squares estimator becomes inflated when one or more eigenvalues of the predictor matrix are close to zero. This results in an estimate that may have low probability of being close to the true value of the vector of regression coefficient β.

When large number of variables are available in a study (or experiment), it is often natural to enquire whether they could be replaced by a fewer number of the variable or of their functions without much loss of information for convenience in the presentation, analysis, and interpretation. Mclachlan (1992) pointed out that discarding the last few principal component of the covariance matrix is less likely to throw away valuable information. Principal components

$Z_1, Z_2, \ldots, Z_m$ which are linear functions of the original k variables $X_1, X_2, \ldots, X_k$ are suggested for this purpose. One good thing about PCA is that once you have found pattern in the data, you can compress the data, by reducing the number of dimension in the data, without much loss of information.

In regression analysis, when the number of variables exceeds the number of observations we face singularity problem which makes ordinary least square regression practically impossible. In situation like this supervised principal component regression seems quite an appealing approach because it replaces the k predictor variable with the subset of predictor variables which are selected based on their association with the outcome. Supervised principal component regression (SPCR) is similar to the classical principal component regression (PCR) except that the classical PCR does not establish direct relationship between the response and predictor variables. Frank and Friedman (1993) compared principal component regression, partial least squares, ordinary least squares, ridge regression, and stepwise regression in a simulation study (using average squared prediction error computed over new observation as the criterion), and found that ridge regression came out ahead of the other techniques. Ridge regression is useful when regression is used for parameter estimation or control and does not directly involve prediction. The problem of ridge regression does not provide insight about the subspace of X that explains the response well.

One popular method in PCR is to use the principal components corresponding to the k largest eigenvalues (Frank and Friedman 1993). The problem with this approach is that the magnitude of the eigenvalue depends on X only, and has nothing to do with response variable. Hence it is possible that principal components that relate X to the response are excluded because they may have small eigenvalues. See Jollife (1982) for several real-life examples. Conversely, the approach may include principal component that are unrelated to the response. Consequently the method does not generally pick the most important column space of X that explains the response well. As an alternative approach is the principal components that have the highest correlation with the response, which makes intuitive sense. However, there are criticisms. See Mason and Gunt (1985), and Almoy (1996). In particular Almoy's numerical studies showed that this alternative approach work slightly worse than the component with the largest eigenvalues in the prediction context. Bair et al (2004) described a technique they called supervised principal components that uses a subset of the predictors that are selected based on their association with the outcome. With supervised approach we can extracts information about important predictors from both the relationship between Y, and $X_1, X_2, \ldots, X_k$, and the correlated predictors themselves. The approach computes the first (or first few) principal component of reduced data matrix consisting of only those X variables whose univariate coefficients with response Y exceed a certain a certain threshold ɵ. These principal component(s) are then use to build regression model which can be used to predict outcome. According to Bair and his group this method compares favorably to other techniques. In this study supervised principal component will be compared with other variable selection regression methods with aim of coming up with model that has highest predictive power.

This paper would be organized in six Sections. Section one contains the introduction while the description of the supervised principal component regression procedure is contained in section two. Section three describes the data used in the analysis. Summary and discussion of the result would be contained in Section four. Conclusion would be in Section five while references would be contained in Section six.

2.0 Supervised Principal Component

Let us assume that there are k predictor variables measured on N observations. And let X be an N x k matrix of predictor variables, and y the N x 1 vector of response measurements with outcomes in metric form. Here in nutshell is the description of supervised principal component regression.

1. Compute (univariate) regression coefficient for each predictor variable.
2. Form reduced data matrix consisting of only those features whose univariate coefficient exceeds a certain threshold ω in absolute value (ω is estimated by cross validation)
3. Compute the first (or first few) principal component of the reduced data matrix.
4. Use the principal component as a predictor variable to compute simple regression with the original response variable.
5. Assess the contribution of the k predictor variables using result of (4).
6. Build a reduced model using only the selected important variables.

Assumed that the columns of X (variables) have been centered to have mean zero, the singular value decomposition (SVD) of X is

$$X = ZDV' \quad (1)$$

Z, D, and V are of dimensions Nxm, mxm, and mxk respectively. D is a diagonal matrix of eigenvalues of X, the columns of Z are the principal components, $Z_1, Z_2, \ldots, Z_{m,}$; these are ordered so that

$d_1 \geq d_2 \geq \cdots \geq d_m \geq 0$; and V is the matrix of the eigenvectors of X.

Let $\tau_i$be the k vector of standardized regression coefficients for measuring the univariate effect of each of the response y.

$$\tau_i = \frac{x_j{}' y_j}{\|x\|^p} \quad (2)$$

At p = 2, τ_i is synonymous to $\beta = (X'X)^{-1}X'Y$ of the ordinary least squares method. Actually a scale estimate $\hat{\sigma}$ is missing in each of the τ_i, but since it is common to all, we can omit it. Let C_ω the matrix of the collection of X indices such that $|\tau_i| > \omega$. We denote by X_ω the matrix consisting of X corresponding to C_ω. The SVD of X_ω is

$$X_\omega = Z_\omega D_\omega V_\omega{}' \quad (3)$$

ω is a cutoff value of $|\tau_i|$, and X_ω are only those predictor variables whose coefficients $|\tau_i|$ exceed the cutoff ω. Letting the transformed data $Z_\omega = (z_{\omega,1}, z_{\omega,2}, \ldots, z_{\omega,m})$, we call $z_{\omega,1}$ the first supervised principal component of X matrix, and so on. We now fit a univariate linear model with response y and predictor $z_{\omega,1}$.

$$\hat{y}^{spc,w} = \alpha + \hat{\eta}\, z_{\omega,1} \tag{4}$$

Note that since $z_{\omega,1}$ is a left singular vector of X_ω, it has mean zero and unit norm. Hence $\hat{\eta} = z_{\omega,1}{}'y$, and the intercept α is the mean of y. We use the cross-validation to estimate the best value of ω. In this paper only the first principal component $z_{\omega,1}$ is considered.

Note that from (3)

$$\begin{aligned} Z_\omega &= X_\omega V_\omega' D_\omega{}^{-1} \\ &= X_\omega U_\omega \end{aligned} \tag{5}$$

So, for example $z_{\omega,1}$ is a linear combination of the column of X_ω; $z_{\omega,1} = X_\omega u_{\omega,1}$. Hence our linear regression model estimate can be viewed as a restricted linear model estimate using all the predictor variables in X_ω.

$$\begin{aligned} \hat{y}^{spc,w} &= \alpha + \hat{\eta}\, z_{\omega,1}\, u_{\omega,1} \\ &= \alpha + X_\omega \hat{\beta}_\omega \end{aligned} \tag{6}$$

Where $\hat{\beta}_\omega = \hat{\eta} u_{\omega,1}$. In fact, by padding $u_{\omega,1}$ with zero corresponding to those variables excluded by C_ω our estimate is a linear in all k predictor variables

Given a test vector x^* we can make predictions from our regression model as follows:

1. We center each component of x^* using the means we derived on the training data;

   $x^*_j \leftarrow x^*_j - \bar{x}_j$.

2. $\hat{y}^* = \alpha + x^{*'}_\omega u_{\omega,1} = \alpha + x^{*'}_\omega \hat{\beta}_\omega$

where x^*_ω is the appropriate sub-vector of x^*_ω

In the case of uncorrelated predictors, it is easy to verify that the supervised principal components procedure has the desired behavior: it yields all predictors whose standardized univariate coefficients exceed ω in absolute term.

2.12 **Important score and a reduced predictor**

Having derived the predictor $z_{\omega,1}$ how do we assess the contributions of the p individual features? It is not true that the features that passed the screen $|\tau_i| > 0$ are necessarily important or are the only important features. Instead, we compute the important score as the correlation between each feature and $z_{\omega,1}$

$$imp = cor(x_j, z_{\omega,1}) \quad (7)$$

Feature j with large values of $|imp|$ contributes most to the prediction of y.

Typically all p predictor variables will have non-zero important scores.

The ability of supervised principal component to build a model based on only a small number of inputs is very important for practical applications. For example, a response variable with highly correlated predictor variables will produce unsatisfactory results.

(Bair et al 2004)

3.0 Materials and methods

In order to assess the performance of supervised principal component regression method, two simulated data sets with correlated predictors and a real life data where the number of observations are smaller than the number of predictors were used.

For simulated data the sample sizes of 50 and 100 with the numbers of predictors of 7 and 5 respectively were generated. The real data used were collected from the Nigeria National Bureau of Statistics Publication 2010 .The data is on gross domestic product and the factors affecting it. Among the factors we studied are agriculture(X_1), mining and quarrying(X_2), manufacturing(X_3), public utility(X_4), building and construction(X_5), transportation(X_6), telecommunication(X_7), wholesale and retail trade(X_8), hotel and restaurants(X_9), finance and insurance(X_{10}), real estate and business services(X_{11}), community social and personal services(X_{12}), and producers of government services(X_{13}). The data covered a period of 11 years.

4.0 Results, and Discussion

4.1 Result of simulated data

The two generated data sets were used to test the adequacy of the model obtained by SPCR with those from classical principal component regression, stepwise regression and linear regression model built using predictor variables selected by SPCR. With logarithm transform of Akaike information criterion (AIC) and Schwarz information criterion;

$$AIC = \frac{2k}{n} + ln\left(\frac{Rss}{n}\right)$$

$$SIC = \frac{k}{n}\ln(n) + ln\left(\frac{Rss}{n}\right)$$

(Gujarati 2003)

we obtained the table of residual analysis as

Table 1

Average of the residual result of Simulated Data

	Residual Analysis		
Analytical Method	Average Error	AIC	SIC
SPCR	0.33	28.0102	28.2696
OLS (computed using SPCR selected predictors)	5378.321	23.9038	24.1332
Stepwise Regression	-1191.330	30.1965	30.4259
Classical PCR	384369.3	27.8449	28.0722

Table 1 shows that OLS (computed using SPCR selected predictors) has the minimum estimated error with AIC and SIC values of 23.9038 and 24.1332 respectively, followed by SPCR, classical PCR, and then stepwise regression.
This shows that OLS (computed using SPCR selected predictors) model is more adequate than others. The AIC and SIC values of SPCR and classical PCR are very close to each other, with this one may say that both methods perform equally well.

4.2 Result of real data

The application of SPCR on the real data gave the model for predicting GDP in Nigeria as

$$GDP = 511.0773 + 0.0005X_3 - 0.0001X_6 + 0.0037X_{11} - 0.0056X_{12} - 0.0019X_{13}$$

When the original values of the variables selected SPCR were used in building linear regression model we obtained GDP model as

$$GDP = 1357 - 0.0267X_3 + 0.0322X_6 - 0.6280X_{11} + 0.8660X_{12} + 3.3210X_{13}$$

The classical principal component regression model was obtained as

$$GDP = 0.0017X_1 + 0.0007X_2 + 0.0002X_3 + 0.0001X_4 + 0.0001X_5 + 0.0001X_6 + 0.0001X_7 + 0.0007X_8 + 0.0002X_{10} + 0.0001X_{11}$$

The stepwise regression model using the standardized coefficient was obtained as

$$GDP = -35.540 + 0.432X_1 + 0.930X_2 + 0.066X_4 + 0.106X_6 + 0.050X_7 + 0.032X_{10} + 0.106X_{12} + 0.289X_{13}$$

The presence of X_{13} in three out of the four models and its coefficients indicate that X_{13} is a very strong factor affecting GDP in Nigeria. The four models were used in computing the residual using AIC and SIC. The residual analyses are in the table 2 below;

Table 2

Residual Analysis of Real Data

	Residual Analysis		
Analytical Method	Average Error	AIC	SIC
SPCR	12.3010	12.1542	12.6244
OLS (computed using SPCR selected predictors)	15.3499	8.0253	8.4956
Stepwise Regression	-206257	31.6661	32.1364
Classical PCR	7.6272	7.4235	7.8937

Table 2 shows that classical principal component regression model is the most adequate of all the models studied followed by OLS (computed using SPCR selected predictors), SPCR, and stepwise regression model.

5.0 Conclusion

The results show that stepwise regression performed poorly in both simulated and real data sets with highest value of AIC and SIC. In the simulated data sets OLS (computed using SPCR selected predictors) performed better than other PCR, SPCR, and stepwise regression methods. In the real data set with more number of predictors than observation classical principal component regression outperformed the other methods. These indicate that SPCR and classical PCR competes favourably to stepwise regression when the data is heavily affected by **multicollinearity.**

Reference

Almoy, T. (1996). A simulation study on the comparison of prediction methods when only a few components are relevant. Computational Statistics and Data Analysis. **21,** 87-107.

Bair, B., Hastie, T., Paul, D., and Tibshirani, R. (2004). Prediction by supervised principal

component. Journal of Statistical Association,**101,**119-137.

Frank, I.E., and Friedman, J.H., (1993). A statistical view of some chemometric regression tools. Technometrics, 35, 109-135.

Gujarati, D.N., (2003). Basic Econmetrics. 4th ed. Tata McGraw-Hill, New York.537-538.

Jollife, I. T. (1982). Principal Component Analysis. Springer-Verlag. New York.

Mason, R.L. and Gunst, R.F. (1985). Selected principal components in regression. Stat and Prob. Lett.**3**

McLachlan, G. J. (1992). Discriminant Analysis and Statistical Pattern Recognition. Wiley,

New York.