Nonparametric Principal Components Regression

Barrios, Erniel
University of the Philippines Diliman, School of Statistics, Magsaysay Avenue, Quezon City 1101, Philippines
E-mail: ernielb@yahoo.com; ebbarrios@up.edu.ph

Umali, Jennifer
University of the Philippines Diliman, School of Statistics, Magsaysay Avenue, Quezon City 1101, Philippines

Introduction

In ordinary least squares regression, dimensionality is a sensitive issue. As the number of independent variables approaches the sample size, the least squares algorithm could easily fail, i.e., estimates are not unique or very unstable, (Draper and Smith, 1981). There are several problems usually encountered in modeling high dimensional data, including the difficulty of visualizing the data, slow convergence for models with numerous parameters, bias in variable selection when some important variables are tentatively dropped at some point during the search process, and the problem of multicollinearity that has a number of potential serious effects on the least squares estimates of the regression coefficients (Montgomery and Peck, 1982).

Multicollinearity happens when two or more predictors are highly correlated resulting to undue influence on the quality of estimates of certain parameters. The presence of multicollinearity inflates the standard errors of the parameter estimates, hence, any small change in the data values causes big change in the estimated functional form. It also makes the estimates of the coefficient unreliable (Curto and Pinto, 2007). The dependent variable will further have weaker sensitivity to variation among the independent variables. Thus, multicollinearity makes it difficult to assess the relative importance of predictor variables in the model.

In dealing with high dimensional data in regression analysis, more than the full set of principal components of the independent variables are used as the explanatory variables instead of the original variables. However, since only a subset of the principal components is used in the regression model, there is generally a loss in information resulting to the deterioration of the predictive ability of the function compared to the model generated using the ordinary least squares linear regression using all independent variables (Dunteman, 1989).

Modeling in a high dimensional data often leads to the issue of specification bias and multicollinearity. Instead of the parametric framework in classical regression, the nonparametric regression on the principal component of the independent variables is explored in this paper to mitigate the deterioration of predictive ability of the model while dealing with high dimensionality. The lost information on the excluded predictors due to selection of important variables is compensated by a flexible functional form of the equation. Furthermore, because of the orthogonality of the principal components, PCR can be postulated as an additive model which is comparably easy to interpret than models that accounts for the independent variables simultaneously.

Nonparametric Principal Components Regression

A two-step procedure is proposed to estimate the regression model. The first step chooses the principal components to be included in the model. The second step fits the best smooth function using the kernel estimation. The predictive performance of the model is evaluated through a simulation study.

Suppose that ***x*** can be represented by the principal components ***z*** of the original variables. Then, we fit the model

$$y_i = f_1(z_{i1}) + f_2(z_{i2}) + \ldots + f_k(z_{ik}) + \varepsilon_i \quad \text{for } i=1,2,\ldots,n,\ k<p \qquad (1)$$

Because of the orthogonality of the components z, (1) satisfies the requirements of an additive model, i.e., each term accounts for the individual, non-overlapping contribution of the components $z_1, z_2, .., z_k$. The $z_1, z_2, .., z_k$ components to be included will be chosen based on their contribution in capturing the variance contained in the original set of variables $(x_1, x_2, .., x_p)$

The functions $f_1, f_2, ..,$ and f_k will be estimated through the backfitting algorithm. The component associated with the largest eigenvalue should be entertained first in the algorithm below:

Step 1. Select the smoothing constant *h.* If *h* is set close to 0, $\hat{f}(z)$ converges to the interpolating splines, but when *h* is very large, the function would still exhibit roughness or will simply interpolate the raw data. The smoothing constant is chosen using the cross validation function.

Step 2. Using the Gaussian kernel, compute $\hat{f}(z)$ that minimizes the penalized sum of squares given by $SS^*(h) = \sum_{i=1}^{n}[y_i - f(z_i)]^2 + h\int_{z_{\min}}^{z_{\max}}[f''(t)]^2\,dt$, where z is the component associated with the largest eigenvalue.

Step 3. Given $\hat{f}(z)$, compute the residuals $e_i^* = y_i - \hat{f}(z_i)$.

Step 4. Iterate from Step 2, this time using the component with second largest eigenvalue. In Step 3, the residuals will be computed with the estimates from the first two components. Continue the iteration until all the components to be included are entertained into the model.

The optimality of the backfitting algorithm is established in the literature, see for example, Opsomer (2000), and Mammen, et. al. (1999) for details.

Simulation Study

The proposed procedure is evaluated through a simulation study. The simulation scenarios will consider a fairly general setting for the existence of the multicollinearity problem as well as for high dimensionality. The comparison between the parametric and nonparametric principal components regression will focus on predictive performance only since this is the main issue being addressed in proposing the nonparametric over the parametric model.

There are several factors considered in the simulation study: dimensionality of the data; the level of multicollinearity; the form of the model; and the predictive ability of the model. There are 15 simulation settings and for each of setting, replicate data sets are generated. For datasets with multicollinearity, X_{p+1} is generated as a function of the X_p and $c*\varepsilon_{p,}$ where c is a constant to ensure that the independent variables are collinear but not necessarily a perfect linear relationship. The degree of multicollinearity is adjusted by multiplying the constant c to the error term. The higher the value of c, the weaker the multicollinearity will be. Y is computed as a linear and nonlinear function of the Xs and ε. To adjust the predictive ability of the model, a constant l is again multiplied in the error term. The higher the value of the l, the lower the value of the R^2, thus, the lower the predictive ability of the model. Model fit is simulated to account for possible misspecification usually encountered in actual modeling.

Results and Discussion

When n<p, both the parametric principal component regression (PCR) and nonparametric PCR (NPCR) are compared to ordinary least squares (OLS) that includes all independent variables. On the otherhand, when n>p, only the parametric and nonparametric PCR are compared.

Effect of the Functional Form of the Data-Generating Function

When the dependent variable is a linear combination of the independent variables, the mean absolute percentage error (MAPE) of PCR and NPCR are generally lower than those from the OLS (including all variables, n<p). Furthermore, MAPE for NPCR is smaller than in PCR and OLS regardless of the form of the data-generating model. The smallest MAPE is observed in the data set generated with a nonlinear data-generating model and estimated through NPCR, see Table 1 for details.

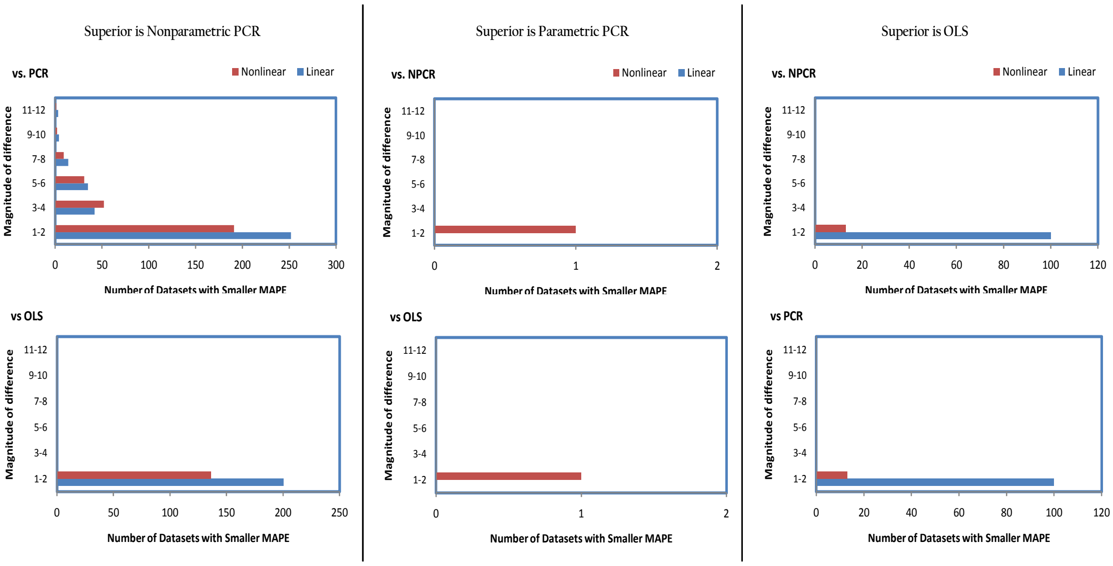
Table 1. Comparison of MAPE from PCR and NPCR by Data-Generating Function

Data Generating Function	Average MAPE			Proportion of with Smallest MAPE		
	OLS	PCR	NPCR	OLS	PCR	NPCR
Linear	42.03%	31.92%	26.55%	22.20%	0.00%	77.80%
Non-linear (Exponential Family)	12.71%	7.59%	5.63%	4.30%	0.30%	95.30%

MAPE is always smaller for datasets with nonlinear combination of the Xs than datasets with linear combination of the Xs except when OLS with all independent variables is used. On the contrary, the highest MAPE are observed when OLS is used in a dataset where the dependent variable is simulated as a linear combination of the independent variables. For both functional forms, there are more datasets with smaller MAPE obtained using the nonparametric PCR than using the parametric PCR and OLS especially when the dependent variable is a nonlinear combination of the Xs.

The magnitude of the difference of the MAPE for the three types of regression analysis is presented in Figure 1. Regardless of the functional form of the data-generating model, in cases where the OLS and parametric PCR are superior over the nonparametric PCR, the MAPE generated using the parametric PCR and OLS is at most twice smaller compared to using the nonparametric PCR. On the other hand, for cases where the nonparametric PCR is better than parametric PCR, the generated MAPE using the former can be up to twelve times smaller for cases where the functional form of the data-generating model is linear and eight times smaller for model where the functional form is nonlinear than the result of the latter.

Figure 1. Magnitude of MAPE by Data-Generating Function



Effect of the Data Dimensionality

For both high and low dimensional datasets, there is a tremendous achievement in model fit for nonparametric PCR compared to the parametric counterpart. The MAPE generated using the nonparametric PCR technique is smaller compared to the parametric PCR (*see Table 2*).

Table 2. Comparison of MAPE from PCR and NPCR by Data Dimensionality

Data Generating Function	Average MAPE			Proportion of with Smallest MAPE		
	OLS	PCR	NPCR	OLS	PCR	NPCR
High Dimensional (n<p)	.	5.02%	1.63%	0.00%	0.00%	100.00%
Low Dimensional (n>p)	32.25%	33.63%	29.22%	25.10%	0.20%	74.70%

For high dimensional data, the smallest MAPE are observed under nonparametric PCR. In fact, the values of the MAPE for a high dimensional data using nonparametric PCR are all below 10% indicating adequacy of the predictive ability of the model. Conversely, the highest MAPE are observed under parametric PCR. Even low dimensional data, the smallest MAPE are also observed under nonparametric PCR. In contrast, the highest MAPE are observed under parametric PCR. In low dimensional data, 25% of the datasets have smaller MAPE using the OLS technique and 75% of datasets with smaller MAPE is exhibited by nonparametric PCR. Even with lost variance due to excluded principal components, nonparametric PCR can still outperform OLS that includes all predictors in the model. When the dimension of the data is high, all simulated datasets that have a smaller MAPE produced using nonparametric PCR can go up to ten times smaller compared to the parametric PCR.

The predictive ability of both PCR and NPCR suffers tremendously when the dimension of the data is low as a result of the discarded principal components associated with very small eigenvalues. In fact, the maximum value of MAPE even reached 80%. Principal components tend to leave out huge amount of information contained in the data. But, nonparametric PCR still exhibits advantage over parametric PCR.

Overall, regardless of the dimension of the data, there are more datasets with smaller MAPE using the nonparametric PCR versus the parametric PCR. Nonparametric PCR is better than parametric PCR for low dimensional data. This advantage is even magnified for high dimensional data.

Effect of Misspecification Error

The effect of model misspecification is simulated by multiplying the error terms with a constant that will consequently affect model fit. If the modeler failed to account all possible sources of explanation for the variation in y, e.g., when the model is misspecified, then OLS will manifest some degree of bias. The effect of model fit is assessed only for low dimensional datasets and the dependent variable is a linear combination of the independent variables.

In Table 3, for datasets without misspecification error, MAPE using the three techniques are almost similar. But still, lowest MAPE is observed when nonparametric PCR is used. On the other hand, for datasets with misspecification errors are at all times slightly larger when parametric PCR or OLS are used instead of nonparametric PCR.

Table 3. Comparison of MAPE from PCR and NPCR by Misspecification Error

Data Generating Function	Average MAPE			Proportion of with Smallest MAPE		
	OLS	PCR	NPCR	OLS	PCR	NPCR
No Misspecification Error	4.24%	4.92%	4.31%	54.00%	0.00%	46.00%
With Misspecification Error	79.82%	82.44%	72.87%	12.70%	0.00%	87.30%

Conclusions

Principal components regression is used to address the problems brought about by multicollinearity usually associated with high dimensional data. The method usually selects the most important components and as a result, the variance contribution of other variables is left out resulting to bias in the estimated model. This problem is mitigated through the specification of the model in a nonparametric context.

Nonparametric principal components regression is generally better than the parametric principal components regression when the true functional form of the model that generates the data is not linear. As the independent variables outnumbered the sample size, the predictive ability of the proposed procedure becomes more superior compared to the usual parametric principal components regression.

For low dimensional datasets and without misspecification error, the predictive ability of the nonparametric principal components regression and the parametric principal components regression is fairly similar. However, when the data-generating model has misspecification error, the nonparametric principal components regression technique significantly improves the predictive ability of the model compared to the parametric principal components regression.

In general, the proposed nonparametric principal components regression procedure can generate a model that effectively resolves the problem of multicollinearity as well as high dimensionality. While the proposed nonparametric PCR is generally advantageous over parametric PCR in a wide variety of cases, the advantage is maximized when data is high dimensional and the functional form of the data-generating model is nonlinear.

REFERENCES (RÉFERENCES)

Curto, J. and Pinto, J. (2007), New multicollinearity indicators in linear regression models. International Statistical Review, 75(1):114-121.

Draper, N. and Smith, H. (1981), Applied Regression Analysis, second edition. New York: John Wiley.

Dunteman, J. (1989), Principal Component Analysis. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-069. Thousand Oaks, CA: Sage.

Mammen, E., Linton, O., and Nielsen, J. (1999), The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. The Annals of Statistics, 27(5): 1443-1490.

Montgomery, D. and Peck, E. (1982), Introduction to Linear Regression. New York: Wiley.

Opsomer, J. (2000), Asymptotic Properties of Backfitting Estimators. Journal of Multivariate Analysis, 73: 166-179.

RÉSUMÉ (ABSTRACT)

Modeling of high dimensional data is often impaired with specification bias and multicollinearity. Principal components regression can resolve multicollinearity but specification bias remains due to the selection only of the important principal components to be included in the model, further resulting to the deterioration of predictive ability of the model. We propose the principal components regression in a nonparametric framework to address the multicollinearity problem (and high dimensionality of predictors) while minimizing (or possibly eliminating) the specification bias that affect predictive ability of the model. The simulation study illustrated how the proposed nonparametric principal components regression address the multicollinearity problem and resolve the issue of high dimensionality while retaining higher predictive ability relative to parametric principal components regression model.