12-12-2018

The Principal Problem with Principal Components Regression

Heidi Margaret Artigue
Pomona College

Heidi Margaret Artigue

The Principal Problem with Principal Components Regression

Heidi Artigue
Pomona College

Gary Smith
Pomona College

Abstract

Principal components regression (PCR) reduces a large number of explanatory variables down to a small number of principal components. PCR is thought to be more useful, the more numerous the potential explanatory variables. The reality is that a large number of candidate explanatory variables does not make PCR more valuable; instead, it magnifies the failings of PCR.

Corresponding author:
Gary Smith
Department of Economics
Pomona College
425 N. College Avenue
Claremont CA 91711
gsmith@pomona.edu

running title: Principal Components Regression
keywords: principal components regression, PCA, factor analysis, Big Data, data reduction
word count: 4,651

The Principal Problem with Principal Components Regression

Pearson (1901) and Hotelling (1933, 1936) independently developed principal component analysis, a statistical procedure that creates an orthogonal set of linear combinations of the variables in an n x m data set X via a singular value decomposition,

$$X = U\Sigma V'$$

where U is an n x m matrix with orthonormal columns, Σ is an m x m diagonal matrix with the ordered singular values, and V is an m x m orthonormal matrix. The non-negative eigenvalues of $X'X$ are the squared diagonal elements of Σ, the eigenvectors of $X'X$ are the columns of V, and the principal components of X are given by XV.

Hotelling (1957) and Kendall (1957) recommended replacing the original explanatory variables in a multiple regression model with their principal components. This replacement evolved into a recommendation by several prominent statisticians that components with small variances can be safely omitted from a regression model (Hocking 1976, Mansfield, Webster, and Gunst 1977, and Mosteller and Tukey 1977). Thus, principal components regression (PCR) discards the eigenvectors that have the smallest eigenvalues, in contrast to other procedures like surrogate regression (Jensen and Ramirez 2010) and raise regression (Garcia, Garcia, and Soto 2011) that increase the magnitude of the small eigenvalues.

PCR enthusiasts evidently believe that components with small variances are of little use in predicting variations in the dependent variable. Mansfield, Webster, and Gunst explicitly state that, “The small magnitude of the latent root indicates that the data contain very little information on the predictiveness of those linear combinations (page 38)”. Mosteller and Tukey argued that,

> A malicious person who knew our x’s and our plan for them could always invent a y to

> make our choices look horrible. But we don't believe nature works that way—more nearly that nature is, as Einstein put it (in German), "tricky, but not downright mean." (pp. 397-398)

Hadi and Ling (1998) show by theory and example that PCR may discard a principal component that is perfectly correlated with the variable being predicted, while retaining components that are completely uncorrelated with the dependent variable. Our point is more general. The principal problem with principal components regression is that it imposes constraints on the coefficients of the underlying independent variables that have nothing whatsoever to do with how these variables affect the dependent variable in the regression model.

Hadi and Ling note that PCR advocates argue that, "Because the PCs . . . are orthogonal, the problem of multicollinearity disappears completely, and no matter how many PCs are actually used, the regression equation will always contain all of the variables in X (because each PC is a linear combination of the variables in X." The problem we highlight is that, while all of the original explanatory variables may be retained, their estimated coefficients are distorted by PCR in ways that diminish the accuracy of the model when it used to make predictions with fresh data.

Principal components regression (PCR) is now commonplace. A principal components transformation of the original explanatory variables is used to create a set of orthogonal eigenvectors, with the corresponding eigenvalues representing the fraction of the variance in the original data that is captured by each eigenvector. The principal components selected for the multiple regression model are then based on a rule such as the largest eigenvalues that capture at least 80 percent of the total variance. A few examples from a wide variety of fields are Cowe and

McNicol (1985), Stock and Watson (2002), Price, Patterson, Plenge, Weinblatt, Shadick, and Reich (2006), Dray (2008), Sanguansat (2012), Sainani (2014), Qi and Roe (2015), and Sabharwal and Anjum (2016).

Some argue that PCR solves the multicollinearity problem created by high correlations among the original explanatory variables; for example, Kudyba (2014), Alibuhtto and Peiris (2015). However, a transformation that retains all the principal components doesn't affect the implicit estimated values or standard errors of the coefficients of the original variables or the predicted values of the dependent variable. The regression model *is* affected if some of the principal components are omitted, but, as will be illustrated later, this is because restrictions with no theoretical basis are imposed on the original parameters.

More recently, PCR has become popular in exploratory data analysis where there is a dauntingly large number of candidate explanatory variables and the researcher wants to let the data determine the final model; for example, Sakr and Gaber (2014), Taylor and Tibshirani (2015), Jolliffe and Cadima (2016), Verhoef, Kooge, and Walk (2016), George, Osinga, Lavie, and Scott (2016), Chen, Zhang, Petersen, and Müller (2017).

Among others, Gimenez and Giusanni (2017) emphasize that it is difficult to interpret the coefficients of the principal components because they are weighted averages of the coefficients of the underlying explanatory variables. Others criticize PCR for its linearity and propose a variety of nonlinear weighting schemes; for example, Liu, Li, McAfee, and Deng (2012), Deng, Tian, and Chen (2013), Yuan et al. (2015), Bitetto, Mangone, Mining, and Giannossa (2016), and Yu and Khan (2017).

These issues are not the most serious problem with principal components regression. The

eigenvector weights depend solely on the correlations among the explanatory variables, with no regard for the dependent variable that the model will be used to predict. As a consequence, PCR may constrain the coefficients of the original explanatory variables in ways that cause the model to fare poorly with fresh data. Specifically, the constraints that the eigenvector weights impose on the implicit estimates may cause the estimated coefficients of nuisance variables to be large, while the estimated coefficients of important explanatory variables may be very small or have the wrong sign.

The Appendix uses a very simple model to provide a detailed example of the practice and pitfalls of principal components regression. We also use a Monte Carlo simulation model to demonstrate how this core problem with principal components regression is exacerbated in large data sets.

A Simulation Model

All the explanatory variables in our Monte Carlo simulations were generated independently in order to focus on the fact that a principal components analysis might be fooled by purely coincidental, temporary correlations among the candidate explanatory variables, some of which are nuisance variables that are independent of the true explanatory variables and of the variable being predicted, and might be useless, or worse, out-of-sample.

Two hundred observations for each candidate explanatory variable were determined by a Gaussian random walk process:

$$X_{i,t} = X_{i,t-1} + \varepsilon_{i,t} \qquad \varepsilon \sim \mathrm{N}[0, \sigma_{\mathrm{x}}] \tag{1}$$

where the initial value of each explanatory variable was zero, and ε was normally distributed with mean 0 and standard deviation σ_{x}. The central question is how effective principal

components regression is at estimating models that can be used to make reliable predictions with fresh data. So, in each simulation, 100 observations were used to estimate the model's coefficients, and the remaining 100 observations were used to test the model's reliability.

All of the data were centered by subtracting the sample means. The in-sample data were centered on the in-sample means and the out-of-sample data were centered on the out-of-sample means so that the out-of-sample predictions would not be inflated if the in-sample and out-of-sample means differed.

Five randomly selected explanatory variables (the *true* variables) were used to determine the values of a dependent variable

$$Y_t = \sum_{i=1}^{5} \beta_i X_{i,t} + \nu_t, \qquad \upsilon \sim \mathrm{N}[0, \sigma_y] \tag{2}$$

where the value of each β coefficient was randomly determined from a uniform distribution ranging from 2 to 4, and υ is normally distributed with mean 0 and standard deviation σ_y. The range 0 to 2 was excluded because the real variables presumably have substantial effects on the dependent variable. Negative values were excluded so that we can compare the average value of the estimated coefficients to the true values. The other candidate variables are *nuisance* variables that have no effect on Y, but might be coincidentally correlated with Y.

A principal components analysis was applied to the in-sample data to determine the eigenvalues, eigenvectors, and principal components. The multiple regression model was estimated by using the principal components associated with the largest eigenvalues such that at least 80 percent of the variation in the explanatory variables is explained by these components.

Our base case was $\sigma_x = 5$, $\sigma_y = 20$, and 100 candidate variables, but we also considered all

combinations of $\sigma_x = 5$, 10, or 20; $\sigma_y = 10$, 20, or 30; and 10, 50, 100, 500, or 1000 candidate variables. One million simulations were done for each parameterization of the model.

Results

The number of principal components included in a multiple regression equation is not affected by the standard deviation of *Y* since the eigenvalues do not depend on *Y*, just the correlations among the candidate explanatory variables. For the same reason, the number of included principal components does not depend on whether the candidate variables truly affect the dependent variable or are merely nuisance variables.

In our simulations, it also turned out that the assumed standard deviation of the explanatory variable hardly mattered either, at least for the range of values considered here; so, we only report the results for our base case of $\sigma_x = 5$ and $\sigma_y = 20$.

With 100 candidate variables, the average PCR equation had 3.01 principal components. Table 1 shows that the average number of components retained increased with the number of candidate variables.

We used the estimated coefficients of the principal components included in the multiple regression model to calculate the implicit estimates of the coefficients of the five real variables and each of the nuisance variables. The expected value of the coefficient of each of the five real variables is 3.0; the true coefficient of each nuisance variable is 0.

Table 1 shows that the average value of the implicit estimated coefficients of the nuisance variables was close to zero, while the average value of the implicit estimates of the coefficients of the true explanatory variables was substantially less than 3 and approached zero as the number of candidate variables increased. This reflects our earlier comment that the construction of

principal components using eigenvector weights imposes unwelcome constraints on the estimated coefficients of the explanatory variables. As the number of candidate variables increases, they become essentially indistinguishable, with estimates that average near zero and consequently do not capture the importance of the real explanatory variables that determine the dependent variable. As the coefficient estimates become essentially noisy, the model becomes less useful for making predictions.

Table 2 uses three metrics to compare the in-sample and out-of-sample prediction errors. The first is the simple correlation between the actual and predicted value of the dependent variable. The second metric is the mean absolute error (MAE)

$$\text{MAE} = \frac{\sum_{t=1}^{n} \left| \hat{Y} - Y \right|}{n} \tag{3}$$

The third metric is the root mean square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n} \left(\hat{Y} - Y \right)^2}{n}} \tag{4}$$

The first row, “5M” in Table 2, is a baseline, using multiple regression estimates with the five true explanatory variables. The other estimates use the principal components with the largest eigenvalues. The principal components models consistently performed far worse out-of-sample than in-sample. As the number of candidate variables increased, the in-sample fit worsened somewhat, while the out-of-sample fit deteriorated substantially.

The results are robust with respect to the number of observations. An increase in the number of observations improves the precision of the estimated coefficients of the principal components, but does materially affect the results, because the flaw in PCR is that correlations among the

explanatory variables are used to constrain the implicit estimates of the model's original coefficients and, on average, these correlations are not affected by an increase in the number of observations. For example, with $\sigma_x = 5$, $\sigma_y = 20$, and 100 candidate variables, 100,000 simulations with 1,000 observations gave results that were essentially the same as in the case of 100 observations: 2.989 versus 3.00 average number of included components; 0.0951 versus 0.091 average estimated coefficients of the true variables; 0.0001407 versus 0.00004 average estimated coefficients of the nuisance variables; and 0.8432 versus 0.819 in-sample and 0.1423 versus 0.144 out-of-sample average correlations between the predicted and actual values of the dependent variable.

The conclusions are also little affected by in-sample correlations among the explanatory variables. We focused on independent candidate variables because we wanted to emphasize the reality that PCR will often give large weights to nuisance explanatory variables even if they are independent of the true explanatory variables. For comparison, we also considered the case of candidate variables with 0.9 pairwise correlations. Table 3 shows the results for the base case of 200 observations (half in-sample and half out-of-sample) and 100 candidate variables. If the candidate variables happen to be highly correlated in-sample, but uncorrelated out-of-sample, PCR tended to choose fewer components (an average of 1.40 versus 3.00), have roughly equal small coefficients for all the variables, and have a better fit in-sample with an equally poor fit out-of-sample. The weaknesses of PCR evidently do not hinge on the in-sample correlations among the explanatory variables.

On the other hand, Table 3 also shows that PCR did relatively well if the explanatory variables happen to be highly correlated both in-sample and out-of-sample. In the first two

scenarios shown in Table 3, the independence of the explanatory variables out-of-sample exposed the PCR pitfall of putting inappropriate weights on the explanatory variables. If the explanatory variables happen to continue to be highly correlated out-of-sample, then these inappropriate weights are not as costly because it doesn't matter as much whether the estimation procedure can distinguish between true variables and nuisance variables.

Conclusion

The promise of principal components regression is that it is an efficient way of selecting a relatively small number of explanatory variables from a vast array of possibilities, based on the correlations among the explanatory variables. The problem in that the eigenvector weights on the candidate variables have nothing to do with their relationship to the variable being predicted. Mildly important variables may be given larger weights than important variables. Nuisance variables may be given larger weights than the true explanatory variables. The coefficients of the true explanatory variables may be given the wrongs signs.

It might be thought that the larger the number of possible explanatory variables, the more useful is the data reduction provided by principal components. The reality is that principal components regression is less effective and more likely to be misleading, the larger is the number of potential explanatory variables.

Appendix A Principal Components Regression Example

Equations 1 and 2 were used to generate twenty observations for four explanatory variables, of which two variables, X_1 and X_2, were used with randomly determined coefficients (3.092 and 3.561, respectively) to determine the values of the dependent variable Y. The other two explanatory variables, X_3 and X_4, were nuisance variables. To keep the standard errors comparable to the main paper, we used $\sigma_x = 5$ and $\sigma_y = 5$. The first ten observations were used for the in-sample statistical analysis, with the ten remaining observations reserved for an out-of-sample test of the model. These data are shown in Table A1.

The eigenvectors and eigenvalues for the four explanatory variables are shown in Table A2. The sum of the eigenvalues is 1,778.42, with the first and second eigenvalues a fraction 0.601 and 0.287 of the total, respectively. Using the 0.80 rule, the two principal components corresponding to these eigenvalues were used in the multiple regression equation.

The first two principal components are

$$PC_1 = 0.7536X_1 - 0.4429X_2 + 0.2586X_3 + 0.4112X_4 \quad (1)$$

$$PC_2 = -0.5423X_1 - 0.0320X_2 + 0.6124X_3 + 0.5743X_4 \quad (2)$$

The absolute values of the weights were larger for the first explanatory variable than for the second, even though the true coefficient of the second variable was larger than the true coefficient of the first variable (3.092 versus 3.561). The weights given the two nuisance variables were comparable to the weights given the real variables. Notice also, that in the first principal component, the weights for the first and second explanatory variables have opposite signs, even though their true coefficients have the same sign. The inescapable problem is that the principal component weights are derived from the correlations among the explanatory variables

with no concern for how the dependent variable is related to the explanatory variables.

If we had used only the first component in our regression model, the implicit coefficients of X_1 and X_2 would necessarily have had opposite signs (one would have an incorrect sign) and the implicit coefficients of the nuisance variables would be substantial. Matters are more complicated when more than one principal component is included in the multiple regression equation, but it remains true that the implicit estimates of the coefficients of the original explanatory variables are constrained by the principal component weights—which depend on the correlations among the explanatory variables rather than their effects on the dependent variable.

The matrix multiplication of the original data by the eigenvector weights gives the principal components shown in Table A3. Using the 0.80 rule, a multiple regression using the first two principal components gave these estimates, with the standard errors shown in parentheses

$$\begin{array}{rcccccc} Y & = & 0.000 & + & 0.777PC_1 & - & 2.028PC_2 \\ & & (6.399) & & (0.619) & & (0.895) \end{array} \qquad (3)$$

The substitution of Equations 1 and 2 into the multiple regression Equation 3 gives the implicit estimates of the coefficients of the original explanatory variables shown in Table A4. The coefficient of X_2, the variable with the largest true coefficient, has the wrong sign, and the coefficient of two nuisance variables are substantial.

Equation 3 was used to make out-of-sample predictions for observations 11 through 20. Table A5 shows that the out-of-sample prediction errors were much larger than the in-sample errors, no doubt because the model's estimated coefficients were so inaccurate. For comparison, a naive model that completely ignores the explanatory variables and simply predicts that Y will equal its average value (0) has a MAE of 31.30 and a RMSE of 36.15. The principal components

regression model was somewhat worse than useless for making predictions.

Table A1 Original Data

observation	Y	X_1	X_2	X_3	X_4
1	–19.760	5.134	–10.697	8.379	10.619
2	24.903	6.741	1.246	7.035	10.937
3	–13.865	1.782	–4.638	3.937	3.488
4	–11.974	–2.652	0.713	0.849	1.836
5	–23.935	–9.772	2.985	3.411	–2.995
6	–28.906	–14.161	4.479	–1.394	–6.317
7	–8.628	–8.842	4.113	–6.448	–5.164
8	40.070	0.030	10.750	–5.308	–1.883
9	8.939	4.444	–3.188	–6.411	–6.080
10	33.155	17.297	–5.764	–4.050	–4.440
11	16.855	–2.358	6.622	0.342	14.457
12	32.368	4.027	4.537	2.522	9.209
13	57.864	6.049	10.789	4.635	4.640
14	49.416	1.800	12.446	1.011	1.777
15	–2.782	3.933	–1.563	–1.395	2.848
16	–16.983	–1.153	–4.771	1.514	–4.049
17	–24.526	3.101	–9.064	–0.149	0.582
18	–29.090	–2.512	–6.998	0.481	–2.160
19	–22.178	–4.184	–2.512	–3.523	–11.667
20	–60.943	–8.703	–9.485	–5.439	–15.637

Table A2 The Eigenvalues and Eigenvectors

	Eigenvectors			
Eigenvalues	E_1	E_2	E_3	E_4
1068.35	0.7536	–0.5423	0.3126	0.2006
511.1	–0.4429	–0.032	0.867	0.2261
174.99	0.2586	0.6124	–0.0398	0.746
23.98	0.4112	0.5743	0.3859	–0.5934

Table A3 Principal Components

observation	Y	PC_1	PC_2	PC_3	PC_4
2	24.903	10.844	6.894	7.129	0.392
3	–13.865	5.849	3.597	–2.274	0.176
4	–11.974	–1.340	2.990	0.464	–0.826
5	–23.935	–9.036	5.572	–1.759	3.037
6	–28.906	–15.613	3.055	–2.926	0.881
7	–8.628	–12.276	–2.251	–0.934	–2.590
8	40.070	–6.885	–4.693	8.814	–0.406
9	8.939	0.603	–9.726	–3.466	–1.003
10	33.155	12.715	–14.226	–1.143	1.779

Table A4 True and Estimated Coefficients

Explanatory Variable	True Coefficient	Estimated Coefficient
X_1	3.093	1.686
X_2	3.561	–0.279
X_3	0	–1.041
X_4	0	–0.845

Table A5 Prediction Errors

Mean Correlation		Mean Absolute Error		Root Mean Square Error	
In-Sample	Out-Sample	In-Sample	Out-Sample	In-Sample	Out-Sample
0.700	–0.542	13.27	36.87	16.93	40.73

References

Alibuhtto, M. C., and T. S. G. Peiris, 2015, Principal component regression for solving multicollinearity problem, *5th International Symposium*, Southeastern University of Sri Lanka.

Artemiou, Andreas, and Bing Li. 2009. On Principal Components andRegression: A Statistical Explanation of a Natural Phenomenon, *Statistica Sinica* 19: 1557-1565.

Bitetto, Alessandro, Annarosa Mangone, Rosa Maria Mininni, and Lorena C. Giannossa. 2016. A Nonlinear Principal Component Analysis to Study Archeometric Data. *Journal of Chemometrics* 30 (7): 405-15. https://doi.org/10.1002/cem.2807.

Chen, Kehui, Zhang, Xiaoke, Petersen, Alexander, and Hans-Georg Müller. 2017. Quantifying Infinite-Dimensional Data: Functional Data Analysis in Action, *Statistics in Biosciences* 9 (2): 582-604.

Cowe, Ian A., and James W. McNicol, 1985, The Use of Principal Components in the Analysis of Near-Infrared Spectra, *Applied Spectroscopy* 39 (2): 257-266.

Deng, Xiaogang, Xuemin Tian, and Sheng Chen. 2013. Modified Kernel Principal Component Analysis Based on Local Structure Analysis and Its Application to Nonlinear Process Fault Diagnosis. *Chemometrics & Intelligent Laboratory Systems* 127: 195-209. https://doi.org/ 10.1016/j.chemolab.2013.07.001.

Dray, Stéphane. 2008. On the Number of Principal Components: A Test of Dimensionality Based on Measurements of Similarity Between Matrices. *Computational Statistics & Data Analysis* 52: 2228-2237.

Garcia, C.B., Garcia, J. and Soto, J. 2011. The raise method: An alternative procedure to estimate

the parameters in presence of collinearity, *Quality and Quantity* 45: 403-423.

George, Gerard, Osinga, Ernst C., Lavie, Dovev, and Brent A. Scott. 2016, Big Data and Data Science Methods for Management Research: From the Editors. *Academy of Management Journal* 59 (5): 1493-1507.

Gimenez, Yanina and Guido Giussani. 2018. Searching for the core variables in principal components analysis. *Brazilian Journal of Probability and Statistics* 32 (4): 730-754.

Gunst, Richard F., and R . L. Mason. 1980. *Regression Analysis and its Applications: A Data-Oriented Approach*, NewYork: Marcel Dekker.

Hadi, Ali S., and Robert F. Ling. 1998. Some Cautionary Notes on the Use of Principal Components Regression. *The American Statistician* 52 (1): 15-19.

Hocking, Ronald R. 1976. The Analysis and Selection of Variables in Linear Regression, *Biometrics* 32: 1-49.

Hotelling, Harold. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24: 417-441 and 498-520.

Hotelling, Harold. 1936. Relations between two sets of variates. *Biometrika* 28: 321-377.

Hotelling, Harold. 1957. The Relations of the Newer Multivariate Statistical Methods to Factor Analysis, *British Journal of Statistical Psychology*10 (2): 69-79.

Jensen, D.R. and Ramirez, D.E. 2010. Surrogate models in ill-conditioned systems. *Journal of Statistical Planning and Inference* 140: 2069-2077.

Jolliffe, Ian T., and Jorge Cadima. 2016. Principal Component Analysis: A Review and Recent Developments. *Philosophical Transactions of the Royal Society A, Mathematical, Physical, and Engineering Sciences* 374 (2065): 20150202.

Kendall, Maurice G. 1957. *A Course in Multivariate Analysis*, London: Griffin.

Kudyba, Stephan. 2014. *Big Data, Mining, and Analytics: Components of Strategic Decision Making*, New York: Auerbach.

Linting, Mariëlle, and Anita van der Kooij. 2012. Nonlinear Principal Components Analysis With CATPCA: A Tutorial. *Journal of Personality Assessment* 94 (1): 12-25. https://doi.org/10.1080/00223891.2011.627965.

Liu, Xueqin, Kang Li, Marion McAfee, and Jing Deng. 2012. Application of Nonlinear PCA for Fault Detection in Polymer Extrusion Processes. *Neural Computing & Applications* 21 (6): 1141-48. https://doi.org/10.1007/s00521-011-0581-y.

Mansfield, E. R., Webster, J. T., and R. F. Gunst. 1977. An Analytic Variable Selection Technique for Principal Component Regression, *Applied Statistics* 26 (1): 34-40.

Mosteller, F., and J. W. Tukey. 1977. *Data Analysis and Regression: A Second Course in Statistics*. Reading, Mass.: Addison-Wesley.

Pearson, K. 1901. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine* 2 (11): 559-572. doi:10.1080/14786440109462720.

Prather JC, Lobach DF, Goodwin LK, Hales JW, Hage ML, Hammond WE. 1997. Medical data mining: knowledge discovery in a clinical data warehouse. *Proceedings of the AMIA Annual Fall Symposium*: 101-105.

Price, A. L., Patterson, Nick J., Plenge, Robert M., Weinblatt, Michael E., Shadick, Nancy A., and David Reich. 2006. Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies. *Nature Genetics* 38: 904-909.

Qi, Danyi, and Brian E. Roe, 2015, Household Food Waste: Multivariate Regression and

Principal Components Analyses of Awareness and Attitudes among U.S. Consumers, *PLoS ONE* 11(7): e0159250. https://doi.org/10.1371/ journal.pone.0159250

Sabharwal, Chaman Lal, and Bushra Anjum. 2016. Data Reduction and Regression Using Principal Component Analysis in Qualitative Spatial Reasoning and Health Informatics. *Polibits* 53: 31-42.

Sainani, Kristen L. 2014. Introduction to Principal Components Analysis. *PM&R* 6 (3): 275 - 278

Sakr, Sherif, and Mohamed Medhat Gaber, eds., 2014. *Large Scale and Big Data: Processing and Management*, London: CRC Press.

Sanguansat, Parinya, editor, 2012, *Principal Component Analysis – Engineering Applications*, Rijeka, Croatia: InTech.

Stock, James H., and Mark W. Watson. 2002. Forecasting Using Principal Components From a Large Number of Predictors, *Journal of the American Statistical Association* 97 (460): 1167-1179.

Taylor, Jonathan, and Robert J. Tibshirani, 2015. Statistical learning and selective inference. www.pnas.org/cgi/doi/10.1073/pnas.1507583112

Verhoef, Peter C., Kooge, Edwin, and Natasha Walk. 2016, *Creating Value with Big Data Analytics: Making Smarter Marketing Decisions*, Abingdon, UK: Routledge.

Yu, Hongyang, and Faisal Khan. 2017. Improved Latent Variable Models for Nonlinear and Dynamic Process Monitoring. *Chemical Engineering Science* 168: 325-38. https://doi.org/ 10.1016/j.ces.2017.04.048.

Yuan, Xiaofeng, Lingjian Ye, Liang Bao, Zhiqiang Ge, and Zhihuan Song. 2015. Nonlinear Feature Extraction for Soft Sensor Modeling Based on Weighted Probabilistic PCA.

Chemometrics & Intelligent Laboratory Systems 147: 167-75. https://doi.org/10.1016/j.chemolab.2015.08.014.

Table 1 Average Number of Principal Components and Estimated Coefficients $\sigma_x = 5$, $\sigma_y = 20$

Number of Candidate Variables	Average Number of Included Components	Average Estimated Coefficient	
		True Variables	Nuisance Variables
5	2.04	1.224	N/A
10	2.44	0.733	–0.00019
50	2.95	0.177	0.00005
100	3.00	0.091	0.00004
500	3.01	0.018	0.000009
1,000	3.02	0.009	0.0000004

Table 2 In-Sample and Out-of-Sample Prediction Errors, $\sigma_x = 5$, $\sigma_y = 20$

	Mean Correlation		Mean Absolute Error		Root Mean Square Error	
Candidates	In-Sample	Out-Sample	In-Sample	Out-Sample	In-Sample	Out-Sample
5M	0.983	0.980	15.47	17.47	19.34	21.79
5	0.835	0.542	41.29	81.51	51.14	97.42
10	0.825	0.410	44.95	94.08	55.69	112.33
50	0.820	0.200	47.51	105.44	58.83	125.78
100	0.819	0.144	47.66	106.37	59.01	126.96
500	0.818	0.061	48.42	108.76	59.91	129.70
1000	0.817	0.047	48.94	110.77	60.69	132.62

5M: multiple regression with five true variables; the other estimates use principal components

Table 3 One Hundred Highly Correlated Candidate Variables, $\sigma_x = 5$, $\sigma_y = 20$

	Correlation Among Candidate Variables		
	None	In-Sample Only	In- and Out-of-Sample
Average Number of Included Components	3.00	1.43	1.43
Average Estimated Coefficient			
True Variables	0.091	0.158	0.158
Nuisance Variables	0.00004	0.142	0.142
Mean Correlation			
In-Sample	0.819	0.981	0.981
Out-of-Sample	0.144	0.198	0.972
Mean Absolute Error			
In-Sample	47.66	33.41	33.41
Out-of-Sample	106.37	105.88	50.65
Root Mean Square Error			
In-Sample	59.01	41.27	41.27
Out-of-Sample	126.96	126.33	60.89