Artigue & Smith, *Cogent Mathematics & Statistics* (2019), 6: 1622190
https://doi.org/10.1080/25742558.2019.1622190

cogent mathematics & statistics

STATISTICS | RESEARCH ARTICLE

The principal problem with principal components regression

Heidi Artigue[1] and Gary Smith[1*]

Received: 07 April 2018
Accepted: 15 May 2019
First Published: 24 May 2019

*Corresponding author: Gary Smith, Department of Economics, Pomona College, 425 N. College Avenue, Claremont, CA 91711, USA
E-mail: gsmith@pomona.edu

Reviewing editor:
Zudi Lu, Mathematical Sciences, University of Southampton, Southampton, UK

Additional information is available at the end of the article

Abstract: Principal components regression (PCR) reduces a large number of explanatory variables in a regression model 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.

Subjects: Statistical Computing; Statistics & Computing; Statistical Theory & Methods; Statistics for Business, Finance & Economics

Keywords: principal components regression; PCA; factor analysis; Big Data; data reduction

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.

Gary Smith

ABOUT THE AUTHORS

Gary Smith is the Fletcher Jones Professor of Economics at Pomona College. He received his Ph.D. in Economics from Yale University and has won two teaching awards and written (or co-authored) more than 90 academic papers and fourteen books, most recently The AI Delusion (Oxford 2018) and The 9 Pitfalls of Data Science (Oxford 2019). A particular concern is data mining and other procedures (like principal components regression) that are guided by the statistical properties of the data, rather than expert knowledge of the phenomena being modeled.

Heidi Artigue is a Research Assistant at the Federal Reserve Bank of Philadelphia where she supports research on urban economics. She received her Bachelor of Arts from Pomona College in Economics, with minors in Mathematics and Computer Science. While at Pomona, Ms. Artigue was awarded the Brystine Prize for Economic Research in 2018.

PUBLIC INTEREST STATEMENT

Principal components regression (PCR) is a popular procedure for reducing a large number of explanatory variables in a regression model down to a small number of principal components. This procedure is fatally flawed because it imposes constraints on the coefficients of the explanatory variables that have nothing whatsoever to do with how these variables affect the response variable. The presence of a large number of potential explanatory variables is thought to make PCR more valuable, but actually magnifies the failings of PCR.

© 2019 The Author(s). This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.

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, & Gunst, 1977,; Mosteller & Tukey, 1977). Thus, principal components regression (PCR) discards the eigenvectors that have the smallest eigenvalues, in contrast to other procedures like surrogate regression (Jensen & Ramirez, 2010) and raise regression (Garcia, Garcia, & 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 explanatory variables that have nothing whatsoever to do with how these variables affect the dependent variable in the regression model.

Hadi and Ling note the argument of PCR advocates 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 et al. (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 estimates 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.

Among others, Gimenez and Giussani (2018) 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, Ye, Bao, Ge, and Song (2015), Bitetto, Mangone, Mining, and Giannossa (2016), and Yu and Khan (2017).

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).

Athey (2018) argues more generally that, "An advantage of using unsupervised learning to create covariates is that the outcome data is not used at all; thus, concerns about spurious correlation between constructed covariates and the observed outcome are less problematic." This characteristic is, in fact, the most serious problem with principal components regression. It is a flaw, not a feature. 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.

1. 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 both true explanatory variables and 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:

$$X_{i,t} = X_{i,t-1} + \varepsilon_{i,t} \quad \varepsilon \sim N[0, \sigma_X] \tag{1}$$

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 the data were centered by subtracting the sample means.

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

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

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 σ_X = 5, σ_y = 20, and 100 candidate variables, but we also considered all combinations of σ_X = 5, 10, or 20; σ_y= 10, 20, or 30; and 10, 50, 100, 500, or 1000 candidate variables. One hundred thousand simulations were done for each parameterization of the model.

2. 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 turned out that the assumed standard deviation of the explanatory variable hardly mattered, at least for the range of values considered here; so, we only report the results for our base case of σ_X = 5 and σ_y= 20.

With 100 candidate variables, the average PCR equation had 3.02 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 estimated coefficients of the nuisance variables was approximately zero, while the average value of the 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,

Table 1. Average number of principal components and estimated coefficients, σ_X= 5, σ_y = 20

		Average Estimated Coefficient	
Number of Candidate Variables	**Average Number of Included Components**	**True Variables**	**Nuisance Variables**
5	2.04	1.224	N/A
10	2.44	0.733	0.000
50	2.95	0.177	0.000
100	3.00	0.091	0.000
500	3.00	0.018	0.000
1,000	3.00	0.009	0.000

and consequently do not capture the importance of the real explanatory variables that determine the value of the dependent variable. As the coefficient estimates become increasingly 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):

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

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 not 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 σ_X = 5, σ_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: 2.99 versus 3.00 average number of included components; 0.091 versus 0.091 average estimated coefficients of the true variables; and 0.832 versus 0.819 in-sample and 0.150 versus 0.141 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 initially focused on independent candidate variables because we wanted to emphasize the reality that PCR will often give large weights to nuisance explanatory variables

Table 2. In-sample and out-of-sample prediction errors, σ_X= 5, σ_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	20.75	19.34	25.52
5	0.835	0.541	41.29	145.98	51.14	167.99
10	0.825	0.409	44.95	168.52	55.69	193.96
50	0.820	0.198	47.51	187.63	58.83	216.38
100	0.819	0.141	47.82	190.49	59.19	219.51
500	0.818	0.010	47.85	190.85	59.21	220.33
1000	0.817	0.008	47.87	191.17	59.30	220.71

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

Table 3. One hundred highly correlated candidate variables, σ_x= 5, σ_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	0.000	0.142	0.142
Mean Correlation			
In-Sample	0.819	0.981	0.981
Out-of-Sample	0.144	0.201	0.973
Mean Absolute Error			
In-Sample	47.66	33.41	33.41
Out-of-Sample	190.64	248.71	92.93
Root Mean Square Error			
In-Sample	59.01	41.27	41.27
Out-of-Sample	219.92	274.35	106.47

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 1.43 versus 3.00), have comparable (small) estimated coefficients for the variables, and a somewhat better fit in-sample and worse 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.

3. 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 is 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 wrong 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.

Funding
The authors received no direct funding for this research.

Author details
Heidi Artigue
E-mail: heidi.artigue@gmail.com[1]
Gary Smith[1]
E-mail: gsmith@pomona.edu
[1] Department of Economics, Pomona College, Claremont, USA.

Citation information
Cite this article as: The principal problem with principal components regression, Heidi Artigue & Gary Smith, *Cogent Mathematics & Statistics* (2019), 6: 1622190.

References

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

Athey, S. (2018). The impact of machine learning on economics. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), *The economics of artificial intelligence: An agenda* (pp. 507-547). University of Chicago Press.

Bitetto, A., Mangone, A., Mininni, R. M., & Giannossa, L. C. (2016). A nonlinear principal component analysis to study archeometric data. *Journal of Chemometrics*, *30*(7), 405–415. doi:10.1002/cem.2807

Chen, K., Zhang, X., Petersen, A., & Müller., H.-G. (2017). Quantifying infinite-dimensional data: Functional data analysis in action. *Statistics in Biosciences*, *9*(2), 582–604. doi:10.1007/s12561-015-9137-5

Cowe, I. A., & McNicol, J. W. (1985). The use of principal components in the analysis of near-infrared spectra. *Applied Spectroscopy*, *39*(2), 257–266. doi:10.1366/0003702854248944

Deng, X., Tian, X., & Chen, S. (2013). Modified Kernel principal component analysis based on local structure analysis and its application to nonlinear process fault diagnosis. *Chemometrics and Intelligent Laboratory Systems*, *127*, 195–209. doi:10.1016/j.chemolab.2013.07.001

Dray, S. (2008). On the number of principal components: A test of dimensionality based on measurements of similarity between matrices. *Computational Statistics and Data Analysis*, *52*, 2228–2237. doi:10.1016/j.csda.2007.07.015

Garcia, C. B., Garcia, J., & Soto, J. (2011). The raise method: An alternative procedure to estimate the parameters in presence of collinearity. *Quality and Quantity*, *45*, 403–423. doi:10.1007/s11135-009-9305-0

George, G., Osinga, E. C., Lavie, D., & Scott, B. A. (2016). Big data and data science methods for management research: From the editors. *Academy of Management Journal*, *59*(5), 1493–1507. doi:10.5465/amj.2016.4005

Gimenez, Y., & Giussani, G. (2018). Searching for the core variables in principal components analysis. *Brazilian Journal of Probability and Statistics*, *32*(4), 730–754. doi:10.1214/17-BJPS361

Hadi, A. S., & Ling, R. F. (1998). Some cautionary notes on the use of principal components regression. *The American Statistician*, *52*(1), 15–19.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, *32*, 1–49. doi:10.2307/2529336

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*, 417–441 and 498–520. doi:10.1037/h0071325

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*, 321–377. doi:10.1093/biomet/28.3-4.321

Hotelling, H. (1957). The relations of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology*, *10*, (2), 69–79. doi:10.1111/j.2044-8317.1957.tb00179.x

Jensen, D. R., & Ramirez, D. E. (2010). Surrogate models in ill-conditioned systems. *Journal of Statistical Planning and Inference*, *140*, 2069–2077. doi:10.1016/j.jspi.2010.02.001

Jolliffe, 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*, *374*(2065), 20150202. doi:10.1098/rsta.2015.0202

Kendall, M. G. (1957). *A course in multivariate analysis*. Griffin: London.

Kudyba, S. (2014). *Big Data, mining, and analytics: Components of strategic decision making*. New York: Auerbach.

Liu, X., Li, K., McAfee, M., & Deng, J. (2012). Application of nonlinear PCA for fault detection in polymer extrusion processes. *Neural Computing and Applications*, *21*(6), 1141–1148. doi:10.1007/s00521-011-0581-y

Mansfield, E. R., Webster, J. T., & Gunst, R. F. (1977). An analytic variable selection technique for principal component regression. *Applied Statistics*, *26*(1), 34–40. doi:10.2307/2346865

Mosteller, F., & Tukey, J. W. (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

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*, 904–909. doi:10.1038/ng1847

Qi, D., & Roe, B. E. (2015). Household food waste: Multivariate regression and principal components analyses of awareness and attitudes among U.S. consumers. *PLoS One*, *11*(7), e0159250. doi:10.1371/journal.pone.0159250

Sabharwal, C. L., & Anjum, B. (2016). Data reduction and regression using principal component analysis in qualitative spatial reasoning and health informatics. *Polibits*, *53*, 31–42. doi:10.17562/PB-53-3

Sainani, K. L. (2014). Introduction to principal components analysis. *PM&R*, *6*(3), 275–278. doi:10.1016/j.pmrj.2014.02.001

Sakr, S., & Gaber, M. M. (Eds.). (2014). *Large scale and Big Data: Processing and management*. London: CRC Press.

Sanguansat, P. (Ed.). (2012). *Principal component analysis – Engineering applications*. Rijeka, Croatia: InTech.

Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, *97*(460), 1167–1179. doi:10.1198/016214502388618960

Taylor, J., & Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, *112*, 7629–7634. doi:10.1073/pnas.1507583112

Verhoef, P. C., Kooge, E., & Walk, N. (2016). *Creating value with Big Data analytics: Making smarter marketing decisions*. Abingdon, UK: Routledge.

Yu, H., & Khan, F. (2017). Improved latent variable models for nonlinear and dynamic process monitoring. *Chemical Engineering Science*, *168*, 325–338. doi:10.1016/j.ces.2017.04.048

Yuan, X., Ye, L., Bao, L., Ge, Z., & Song, Z. (2015). Nonlinear feature extraction for soft sensor modeling based on weighted probabilistic PCA. *Chemometrics & Intelligent Laboratory Systems*, *147*, 167–175. doi:10.1016/j.chemolab.2015.08.014

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 (2.022 and 3.403, 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 σ_X = 5 and σ_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 2,735.44, with the first eigenvalue a fraction 0.889 of the total. Using the 0.80 rule, the principal component corresponding to this eigenvalue was used in the multiple regression equation.

The first principal components is

$$PC_1 = -0.3298X_1 + 0.6054X_2 - 0.6512X_3 + 0.3173X_4 \tag{1}$$

The absolute values of the weights given the two nuisance variables are comparable to the weights given the real variables. Notice, also, that 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 these principal component weights were derived from the correlations among the explanatory variables with no concern for how the dependent variable is related to the explanatory variables. Because of this, the implicit coefficients of the true variables, X_1 and X_2, will necessarily have opposite signs (one will have an incorrect sign) and the implicit coefficients of the nuisance variables will be substantial.

Matters would be more complicated if more than one principal component were included in the multiple regression equation, but it remains true that the implicit estimates of the coefficients of the original explanatory variables would be 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 principal component gave these estimates, with the standard errors shown in parentheses,

$$\begin{array}{ccccc} Y & = & 0.000 & + & 1.504PC_1 \\ & & (4.254) & & (0.273) \end{array} \tag{2}$$

The t-value for the coefficient of the first component is a remarkable 5.51 and R^2 = 0.79, seemingly a highly successful model.

The substitution of Equation 1 into the multiple regression Equation 2 gives the implicit estimates of the coefficients of the original explanatory variables shown in Table A4. The estimated coefficient of X_1 has the wrong sign and the absolute values of the coefficients of the nuisance variables are as large as those for the true variables. Despite the impressive in-sample fit, the estimated model is unlikely to be nearly as successful out-of-sample.

Equation 2 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. Despite the 0.890 in-sample correlation, the out-of-sample correlation between the predicted and actual values of Y was *negative*. The out-of-sample MAE and RMSE were roughly double the in-sample values. 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 15.71 and a RMSE of 13.95. 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	20.899	−9.400	7.654	−16.163	4.014
2	41.320	−7.493	14.967	−11.858	8.426
3	9.584	−4.587	7.990	−10.045	9.638
4	16.632	1.872	3.797	−4.053	3.245
5	9.523	−2.325	6.163	−0.779	−0.193
6	6.874	0.825	1.940	3.796	−6.928
7	−11.663	3.343	−5.947	−0.465	−5.427
8	−30.695	4.341	−9.950	8.108	−0.364
9	−56.387	6.518	−20.660	15.646	−5.637
10	−6.087	6.908	−5.954	15.813	−6.774
11	7.878	6.577	−0.024	15.388	−8.043
12	4.724	−0.406	0.912	6.754	−13.972
13	−9.331	−4.455	−0.810	−0.136	−20.521
14	15.417	3.827	−0.791	1.051	−26.797
15	15.709	8.273	−2.250	−1.755	−21.643
16	12.040	6.598	−1.761	−8.400	−31.408
17	−6.672	3.634	−3.366	−3.504	−38.517
18	17.103	4.746	−1.061	−5.460	−31.746
19	−20.269	1.306	−6.963	−2.750	−23.184
20	−30.335	−7.336	−6.182	−4.434	−22.629

Table A2. The Eigenvalues and Eigenvectors

	Eigenvectors			
Eigenvalues	E_1	E_2	E_3	E_4
2,431.96	−0.3298	0.1122	0.1769	0.9205
174.90	0.6054	0.7774	0.1420	0.0949
98.46	−0.6512	0.4693	0.4605	−0.3791
30.11	0.3173	−0.4035	0.8582	−0.0020

Table A3. Principal components

Observation	Y	PC_1	PC_2	PC_3	PC_4
1	20.899	19.533	−4.309	−4.575	−1.809
2	41.320	21.928	1.830	2.570	−1.000
3	9.584	15.949	−2.907	3.968	0.323
4	16.632	5.349	−0.050	1.789	3.613
5	9.523	4.944	4.243	−0.061	−1.260
6	6.874	−3.768	6.177	−3.776	−0.482
7	−11.663	−6.122	−2.277	−5.125	2.701
8	−30.695	−12.851	−3.296	2.777	−0.021
9	−56.387	−26.634	−5.712	0.587	−1.879
10	−6.087	−18.330	6.300	1.845	−0.186

Table A4. True and estimated coefficients

Explanatory Variable	True Coefficient	Estimated Coefficient
X_1	2.022	−0.496
X_2	3.403	0.910
X_3	0	−0.979
X_4	0	0.477

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
0.890	−0.324	10.61	20.43	12.03	22.99

© 2019 The Author(s). This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.

You are free to:
Share — copy and redistribute the material in any medium or format.
Adapt — remix, transform, and build upon the material for any purpose, even commercially.
The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:
Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made.
You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
No additional restrictions

You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

***Cogent Mathematics & Statistics* (ISSN: 2574-2558) is published by Cogent OA, part of Taylor & Francis Group.**

Publishing with Cogent OA ensures:

- Immediate, universal access to your article on publication
- High visibility and discoverability via the Cogent OA website as well as Taylor & Francis Online
- Download and citation statistics for your article
- Rapid online publication
- Input from, and dialog with, expert editors and editorial boards
- Retention of full copyright of your article
- Guaranteed legacy preservation of your article
- Discounts and waivers for authors in developing regions

Submit your manuscript to a Cogent OA journal at www.CogentOA.com