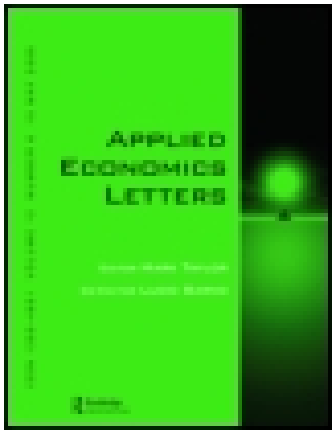

Applied Economics Letters
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rael20

Robust vs. classical principalcomponent analysis in the presence of outliers

Sunil K. Sapra [a]
[a] Department of Economics and Statistics , California State University , Los Angeles, 5151 CA, 90032, United States
Published online: 09 May 2008.

To cite this article: Sunil K. Sapra (2010) Robust vs. classical principalcomponent analysis in the presence of outliers, Applied Economics Letters, 17:6, 519-523, DOI: 10.1080/13504850802046989

To link to this article: http://dx.doi.org/10.1080/13504850802046989

Applied Economics Letters, 2010, **17**, 519–523

Routledge
Taylor & Francis Group

Robust vs. classical principal component analysis in the presence of outliers

Sunil K. Sapra

Department of Economics and Statistics, California State University, Los Angeles, 5151 CA 90032, United States
E-mail: ssapra@calstatela.edu

Principal Component Analysis (PCA) is a very versatile technique for dimension reduction in multivariate data. Classical PCA is very sensitive to outliers and can lead to misleading conclusions in the presence of outliers. This article studies the merits of robust PCA relative to classical PCA when outliers are present. An algorithm due to Filzmoser *et al*. (2006) based on a modification of the projection pursuit algorithm of Croux and Ruiz-Gazen (2005) is used for robust PCA computations for a financial data set as well as simulated data sets. Our simulation results indicate that robust PCA generally leads to greater reduction in model dimension than classical PCA in data sets with outliers.

I. Introduction

Principal components analysis (PCA) is a technique for simplifying data sets by reducing multidimensional data sets to lower dimensions for analysis. It proceeds by computing eigenvalues and eigenvalues of either a correlation or a covariance matrix. PCA is an orthogonal linear transformation of the original data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (the first principal component), the second greatest variance on the second coordinate, and so on. It leads to dimension reduction in a data set while retaining those characteristics of the data set that contribute most to its variance by retaining principal components with largest variances and ignoring those with smallest variances. PCA continues to be the subject of much research, ranging from new model-based approaches to algorithmic ideas from neural networks. It is extremely versatile and has applications in many disciplines including economics, finance and engineering.

Classical PCA uses empirical covariance matrix or correlation matrix of the data (Anderson, 2003) and is very sensitive to outliers. Three robust PCA approaches have been developed to address this difficulty. The first approach due to Croux and Haesbroeck (2000) is based on the eigenvectors of a robust dispersion matrix such as the MCD or an S-estimator and is limited to low-dimensional data. The second approach due to Li and Chen (1985), Croux and Ruiz-Gazen (2005) is based on projection pursuit and can handle high-dimensional data. The third approach (ROBPCA approach) due to Hubert *et al*. (2005) combines projection pursuit with robust scatter matrix estimation and yields more accurate estimates for noncontaminated data and more robust estimates at contaminated data. It is useful for analysis of regression data with both outliers and multicollinearity. Nevertheless, applications of robust PCA to financial and economic data sets are uncommon as are comparisons between robust and nonrobust PCA despite the occasional presence of outliers.

Applied Economics Letters ISSN 1350–4851 print/ISSN 1466–4291 online © 2010 Taylor & Francis
http://www.informaworld.com
DOI: 10.1080/13504850802046989

This article studies the merits of robust PCA relative to classical PCA in the presence of outliers. Our robust PCA is based on the algorithm of Fritz and Filzmoser (2006), which is a modification of projection pursuit algorithm of Croux and Ruiz-Gazen (2005). Section II presents a discussion of the robust PCA algorithm used in the article. Section III presents an application of robust PCA to a financial data set and compares the results with those of the classical PCA. Section IV presents results of a small simulation study for the comparison of classical and robust PCA techniques and Section V provides concluding remarks.

II. Robust PCA Based on the Projection Pursuit Approach

In this section, we discuss the robust PCA based on the projection pursuit (PP) approach of Huber (1985), a method due to Filzmoser *et al.* (2006). The method is an extension of the algorithm of Croux and Ruiz-Gazen (2005), which uses each observation for the construction of candidate directions. The main objective is to find the direction, which maximizes an objective function such as a robust spread measure for robust PCA. The PP-based estimator is based on the following algorithm (Croux and Ruiz-Gazen (2005), pp. 215–216).

Let $X=\{x_1, x_2, \ldots, x_n\}$ be the sample and $\hat{\mu}_n(X)$ a location estimate based on this sample, $x_1, x_2, \ldots, x_n \in R^p$. Let $1 \le q \le p$ be the number of principal components to be computed. Select a scale estimator S_n as projection index.

For $k=1$, set $x_i^1 = x_i - \hat{\mu}_n(X)$ for $i=1,2,\ldots,n$, where $\hat{\mu}_n(X) = \arg\min_{\mu \in R^p} \Sigma_{i=1}^n \|x_i - \mu\|$, $\|\ \|$ is the Euclidean norm. Define $A_{n,1}(X) = \{x_i^1/\|x_i^1\| \; 1 \le i \le n\}$ and set

$$\hat{\delta}_{s_n,1} = \arg \max_{a \in A_{n,1}(X)} S_n(a'x_1^1, a'x_2^1, \ldots, a'x_n^1)$$

Compute the scores on the first component as $y_i^1 = \hat{\delta}_{s_n,1} x_i^1$ for $i=1,2,\ldots,n$.

For $k=2,3,\ldots,q$, define recursively

1. for $i=1,2,\ldots,n$, $x_i^k = x_i^{k-1} - y_i^{k-1}\hat{\delta}_{s_n,k-1}$,
2. the set $A_{n,k}(X) = \{x_i^k/\|x_i^k\|;\ 1 \le i \le n\}$,
3. the estimated eigenvector $\hat{\delta}_{s_n,k} = \arg\max \times S_n(a'x_1^k, a'x_2^k, \ldots, a'x_n^k)$.
4. for $i=1,2,\ldots n$, $y_i^k = \hat{\delta}_{s_n,1} x_i^k$ yielding approximations for the eigenvectors and for the vector of scores on the kth principal component $(y_1^k, y_2^k, \ldots, y_n^k)'$. Approximations for $\hat{\lambda}_{s_n,k}$ for $k=1,2,\ldots,q$ for the eigenvalues and for the covariance matrix $\hat{C}_{s_n}$ are computed as $\hat{\lambda}_{s_n,k} = S_n^2(a^t x_1^k, a^t x_2^k, \ldots, a^t x_n^k$ for $k=1,2,\ldots,q$

$$\hat{C}_{s_n} = \sum_{k=1}^{p} \hat{\lambda}_{s_n,k} \delta_{S_n,k} \delta'_{S_n,k}$$

However, in the algorithm of Filzmoser *et al.* (2006) used in the present study, the directions are randomly generated unlike Croux and Ruiz-Gazen (2005). In the first step, variables are sorted in descending order according to the largest scale. Then the optimization is done in the plane spanned by the first two sorted variables, where the candidate directions are constructed by dividing the unit circle into a regular grid of segments. A second approximation of the projection direction is then found by maximizing in the plane formed by the first and the third sorted variable. The procedure is repeated until the last variable has entered the optimization, which completes the first cycle of the algorithm. In a second cycle each variable is in turn again considered for improving the maximal value of the objective function. The algorithm terminates after a fixed number of cycles or when the improvement is considered to be marginal (see Fritz and Filzmoser (2006) for details).

III. Application of Classical PCA and Robust PCA to Log Monthly Returns

We apply the classical PCA and robust PCA to data on monthly log returns of IBM, Hewlett Packard, Intel Corporation, Merril Lynch and Morgan Stanley Dean Witter from January 1990 to December 1999 taken from Tsay (2005). The returns are in percentages and include dividends. The data set has 120 observations. Tables 1 and 2 present, respectively, the results of classical PCA using empirical covariance matrix and robust PCA using robust estimates of the eigenvalues and eigenvectors of the covariance matrix.

Under classical PCA, the first principal component accounts for 50.61% of the total variation in the data, the first two principal components account for 73.55%, the first three principal components account for 86.38%, the first four principal components account for 95.63% and the first five principal components account for 100% of the total variation in the data.

Under robust PCA, the first principal component accounts for 52.73% of the variation in the data, the first two principal components account for

Table 1. Results of classical PCA applied to log monthly returns

Eigenvalues and eigenvectors of covariance matrix

Variances of principal components: eigenvalues of the covariance matrix
$\lambda_1 = 254.02354,\ \lambda_2 = 115.17240,\ \lambda_3 = 64.36848,\ \lambda_4 = 46.42777,\ \lambda_5 = 21.92835.$

Component coefficients: eigenvectors of the covariance matrix
$e_1 = [-0.2460393, -0.4614131, -0.4091571, -0.5215279, -0.5358747]$
$e_2 = [0.3271473, 0.3597586, 0.5845591, -0.4515628, -0.4668299]$
$e_3 = [-0.58551069, -0.42762931, 0.68284069, 0.08230439, 0.03556665]$
$e_4 = [0.69950387, -0.68722189, 0.15343572, 0.11453886, 0.04193738]$
$e_5 = [0.01763066, -0.05030898, 0.03266817, -0.71007502, 0.70134473]$

Importance of components

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
SD	15.9381160	10.7318404	8.0229969	6.81379299	4.6827721
Proportion of variance	0.5061031	0.2294634	0.1282444	0.09250025	0.0436889
Cumulative proportion	0.5061031	0.7355665	0.8638109	0.95631110	1.0000000

Table 2. Results of robust PCA applied to log monthly returns

Eigenvalues and eigenvectors of robust covariance matrix

Variances of principal components: eigenvalues of the robust covariance matrix
$\lambda_1 = 262.90545,\ \lambda_2 = 111.80068,\ \lambda_3 = 57.61837,\ \lambda_4 = 42.99173,\ \lambda_5 = 23.31989.$

Component coefficients: eigenvectors of the robust covariance matrix
$e_1 = [-0.2772273, -0.5593677, -0.3678021, -0.4794886, -0.4950405]$
$e_2 = [-0.53891431, 0.07824198, -0.64970109, 0.22073009, 0.482303100]$
$e_3 = [0.75353075, 0.09660595, -0.59294993, -0.23125024, 0.13338819]$
$e_4 = [0.24012105, -0.81219613, 0.08101142, 0.41861720, 0.31761002]$
$e_5 = [-0.08514512, -0.10950212, 0.29061597, -0.70189164, 0.63533462]$

Importance of components

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
SD	16.21436	10.57358	7.59068	6.55681	4.82907
Proportion of variance	0.527249	0.224213	0.115552	0.086219	0.046767
Cumulative proportion	0.527249	0.751462	0.867014	0.953233	1.0000000

75.15% of the variation, the first three principal components account for 86.70%, the first four principal components account for 95.32% and the first five principal components account for 100% of all variation in the data. These percentages are close to those under classical PCA since the number of outliers in the data set is small-an average of three outliers per variable.

The percentage of total variance explained by the first component and cumulative percentages for the first four components are slightly higher under robust PCA than under classical PCA-a result demonstrated more clearly through the simulation experiment in the next section. Thus, a more parsimonious model is likely to be chosen under robust PCA in the presence of outliers.

IV. Simulation Experiment

We performed a small simulation experiment to compare the performances of classical and robust PCAs. Previous studies have documented performances of such estimators in terms of mean squared errors of robust and nonrobust estimators of eigenvalues of the population covariance matrix of responses. Unlike these studies, we focus on proportions of variance explained averaged across one hundred replications for each principal component using classical and robust PCAs. One hundred samples of sizes 100 and 500 were generated as follows. For $p = 3$, for each sample, 70% of the observations were generated from $N(\mu, \Sigma)$, $\mu = (10, 8, 3)'$, $\Sigma = \text{diag}(100, 2, 4)$ and 30% of the

Table 3. Comparison of proportions of variance explained using classical and robust PCA: $p = 3$, 30% contamination

Sample size	Component 1	Component 2	Component 3
Classical PCA			
100	0.855601	0.114995	0.029414
500	0.853991	0.117855	0.028154
Robust PCA			
100	0.892669	0.107331	
500	0.899022	0.100978	

Table 4. Comparison of proportions of variance explained using classical and robust PCA: $p = 4$, 30% contamination

Sample size	Component1	Component 2	Component 3	Component 4
Classical PCA				
100	0.790768	0.15593	0.034277	0.019655
500	0.784474	0.159976	0.035653	0.019898
Robust PCA				
100	0.753067	0.246933		
500	0.745761	0.226484		

observations were generated from a multivariate t population with 3 degrees of freedom. Similarly, for $p = 4$, for each sample, 70% of the observations were generated from $N(0, \Sigma)$, $\Sigma = \text{diag}(100, 2, 4, 20)$ and the remaining 30% of the observations were generated from a multivariate t distribution with 7 degrees of freedom. This led to a 30% contamination in each sample. The results for $p = 3$ and $p = 4$ are reported in Tables 3 and 4, respectively.

The results for simulation settings in these tables can be summarized as follows.

1. Under robust PCA, all of the variance is explained by the first two principal components regardless of the value of p, the number of data columns. In contrast, under classical PCA, a small amount of variance can be attributed to the third and fourth principal components: 2% under $p = 3$ and about 5% under $p = 4$. Robust PCA appears to lead to greater parsimony than its nonrobust counterpart in the presence of outliers. Under robust PCA, 99% of the total variation in the data can be explained by retaining the first two principal components with the largest variance, which is much higher than the corresponding percentage under classical PCA.
2. As the sample size increases, a slight reallocation of variance among principal components occurs under classical as well as robust PCA: the variance explained by the leading principal components decreases, while the variance explained by the last components increases. Thus, a large sample size may insure one against incorrect dimension reduction.
3. As p, the dimension of the data set increases from 3 to 4, the variance explained by the first principal component decreases substantially while the variance explained by each of the remaining components increases under both classical and robust PCA. Increasing the data dimension appears to lead to a loss of parsimony.

V. Conclusions

This article studied the merits of robust vs. classical PCA. These techniques were applied to a financial data set and the results compared by comparing the variance explained by each component. A small simulation study was also presented. The results suggest that robust PCA generally leads to more parsimonious models than its nonrobust counterpart in the presence of outliers. Applications of robust PCA to large-scale simultaneous equation models in undersized samples containing outliers is of particular interest since parsimony is an important goal due to the large number of variables in these models. The projection pursuit approach works even for data sets with more variables than observations.

References

Anderson, T. W. (2003) *An Introduction to Multivariate Analysis*, Wiley, New York.

Croux, C. and Haesbroeck, G. (2000) Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies, *Biometrika*, **87**, 603–18.

Croux, C. and Ruiz-Gazen, A. (2005) High breakdown estimators for principal components: the projection-pursuit approach revisited, *Journal of Multivariate Analysis*, **95**, 206–26.

Filzmoser, P., Serneels, S., Croux, C. and Van Espen, P. J. (2006) Robust multivariate methods: the projection pursuit approach, in *From Data and Information Analysis to Knowledge Engineering* (Eds) M. Spiliopoulou, R. Kruse, A. Nurnberger, C. Borgelt, and W. Gaul, Springer-Verlag, Heidelberg-Berlin, pp. 270–7.

Fritz, H. and Filzmoser, P. (2006) Robust Principal Component Analysis by Projection Pursuit available at http://www.r-project.org/user-2006/Abstracts/Filzmoser±Fritz.pdf

Huber, P. J. (1985) Projection pursuit, *The Annals of Statistics*, **13**, 435–75.

Hubert, M., Rousseeuw, P. J. and Branden, K. V. (2005) ROBPCA: a new approach to robust principal component analysis, *Technometrics*, **47**, 64–79.

Li, G. and Chen, Z. (1985) Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo, *Journal of the American Statistical Association*, **80**, 759–66.

Tsay, R. (2005) *Analysis of Financial Time Series*, Wiley, New York.