The Korean Communications in Statistics
Vol. 13 No. 2, 2006, pp. 233-241

# LMS and LTS-type Alternatives to Classical Principal Component Analysis

Myung-Hoe Huh [1] and Yonggoo Lee [2]

## Abstract

Classical principal component analysis (PCA) can be formulated as finding the linear subspace that best accommodates multidimensional data points in the sense that the sum of squared residual distances is minimized. As alternatives to such LS (least squares) fitting approach, we produce LMS (least median of squares) and LTS (least trimmed squares)-type PCA by minimizing the median of squared residual distances and the trimmed sum of squares, in a similar fashion to Rousseeuw (1984)'s alternative approaches to LS linear regression. Proposed methods adopt the data-driven optimization algorithm of Croux and Ruiz-Gazen (1996, 2005) that is conceptually simple and computationally practical. Numerical examples are given.

*Keywords* : Principal component analysis (PCA); Projection pursuit; Least squares (LS); Least median of squares (LMS); Least trimmed squares (LTS).

## 1. Background and Aim of Study

Suppose that there are $n$ data points $x_1, \cdots, x_n$, each of which consists of $p$ ($\geqq 3$) inter-correlated continuous measurements. Classical principal component analysis (PCA) portrays high dimensional continuous data points on the lower dimensional space. Because of its data visualizing function, PCA is favored by many statisticians as major exploratory data analysis (EDA) tool.

As EDA tool, PCA can be formulated as follows (Lebart et al. 1984, Huh 1999). Let $v$ be a $p \times 1$ unit vector and consider linear projections of $x_1, \cdots, x_n$ onto $v$, so that the projections provide one-dimensional display of $p$-dimensional objects. In doing so, it is desirable to find a unit vector $v$ such that the sum of squared residual distances is minimum. That is,

$$\text{minimize (w.r.t. } v \text{ )} \sum_{i=1}^{n} \| x_i - (v^t x_i) v \|^2 \quad \text{subject to} \quad v^t v = 1. \tag{1}$$

[1] Professor, Dept. of Statistics, Korea University. Anam-dong 5, Sungbuk-Gu, Seoul 136-701, Korea. Correspondence : stat420@korea.ac.kr

[2] Professor, Dept. of Statistics, Chung Ang University.

Along the unit vector $v_1$ that solves (1), $n$ objects are positioned respectively at $v_1^t x_1, \cdots, v_1^t x_n$. Subsequently, we focus on the residual vectors

$$x_i - (v_1^t x_i) v_1, \ i = 1, \cdots, n,$$

which, we denote by $x_1^{(1)}, \cdots, x_n^{(1)}$, substitute for $x_1, \cdots, x_n$. Next, we try to find another unit vector $v$ that

$$\text{minimize (w.r.t. } v\text{)} \sum_{i=1}^{n} \| x_i^{(1)} - (v^t x_i^{(1)}) v \|^2 \quad \text{subject to} \quad v^t v = 1 . \tag{2}$$

On the subspace spanned by $v_1$ that solves (1) and another unit vector $v_2$ that solves (2), $n$ objects are projected at

$$(v_1^t x_1, v_2^t x_1), \cdots, (v_1^t x_n, v_2^t x_n)$$

on the two-dimensional subspace. In that way, the lower dimensional display extends to three or more dimensional subspace.

Quite clearly, least squares (LS) optimization formulated in (1) and (2) lacks robustness. Therefore least median of squares (LMS) and/or least trimmed squares (LTS) optimization is more desirable particularly for exploratory data analysis (EDA) purpose, as in linear regression (Rousseeuw, 1984).

The aim of this study is to build practical LMS and LTS-type principal component analysis (PCA) algorithm for EDA. To avoid computational complications that is inevitable in brute-force projection pursuit, we use data-driven approach to classical PCA of Croux and Ruiz-Gazen (1996, 2005) and give partial justifications in Section 2. Subsequently, we propose LMS and LTS-type PCA algorithms and give a numerical example in Section 3. Finally in Section 4, we conclude the paper with remarks on the scalability of the proposed algorithms and the relationship with other robust PCA techniques.

## 2. Data-driven Approach to LS-type PCA

For PCA, we assume that the $n \times p$ data matrix $X$ with $x_i^t$ as its $i$ th row is (robustly) centered and scaled unless stated otherwise. It is well known that the eigenvectors $v_1$ and $v_2$ of $p \times p$ symmetric matrix $X^t X / (n-1)$ corresponding to the largest and second largest eigenvalues are the solutions of (1), (2) and so on.

We adopt the simple approach of Croux and Ruiz-Gazen (1996, 2005) that meets the same problem. We call it by "Data-driven Least Squares (LS)-type" PCA Algorithm (Others call it by C-R Algorithm):

Step 1: Treat normalized $x_1, \cdots, x_n$ as possible candidates for $v$, evaluate

$$\sum_{i=1}^{n} \| x_i - (x_k^t x_i)/(x_k^t x_k)\, x_k \|^2, \quad \text{for } k = 1, \cdots, n \tag{3}$$

and find the smallest one and corresponding index $k_1$, from which we obtain the first loading vector $v_1^* = x_{k_1} / \| x_{k_1} \|$ .

Step 2: Replace $x_1, \cdots, x_n$ by residual vectors $x_i^{(1)} = x_i - (v_1^{*t} x_i)\, v_1^*$, $i = 1, \cdots, n$, to form the new data matrix and go back to Step 1 to obtain the second index $k_2$ and second loading vector $v_2^* = x_{k_2}^{(1)} / \| x_{k_2}^{(1)} \|$ . As result, one obtains two-dimensional principal component display

$$(v_1^{*t} x_1, v_2^{*t} x_1), \cdots, (v_1^{*t} x_n, v_2^{*t} x_n).$$

Note that $v_1^*$ is orthogonal to $v_2^*$ but that score vectors

$$s_1 = (v_1^{*t} x_1, \cdots, v_1^{*t} x_n) \quad \text{and} \quad s_2 = (v_2^{*t} x_1, \cdots, v_2^{*t} x_n)$$

may not be so.

Step 3: Repeat Step 2 when additional principal component dimension is needed.

As an illustration, we applied the classical PCA and the Data-driven LS-type PCA to National Track Records dataset consisting of 55($=n$) observations and 8($=p$) variables (Johnson and Wichern, 1992, p.393). All variables are inversely transformed to mitigate distributional skewness. See <Figure 1> and <Figure 2> produced by classical PCA and Data-driven LS-type PCA of inverse transformed data. The left-side graph shows observation scores on first two principal axes, while the right-side graph shows the loading coefficients for variables. In these figures, observation points are labeled with "abbreviated Nations" and variable points are labeled with "A"-"G" for 100m, 200m, 400m, 800m, 1.5Km, 5Km, 10Km and "M" for marathon. Differences between two figures are pretty small.

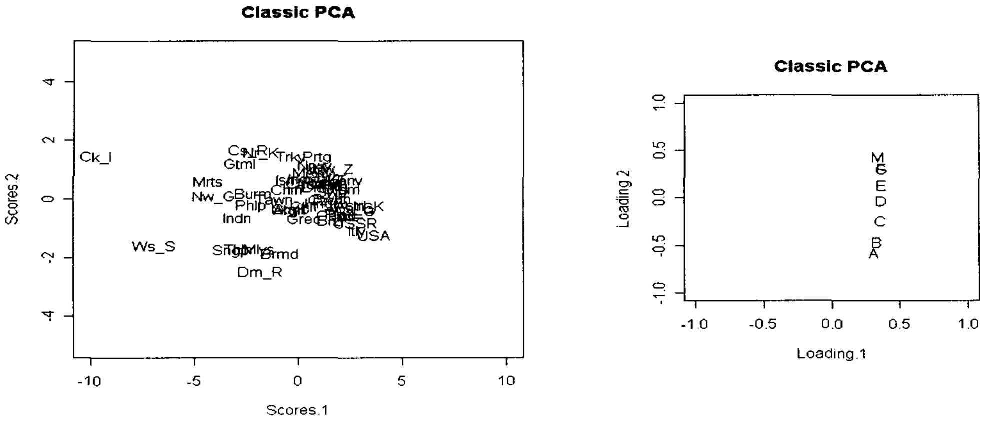
Strictly speaking, Data-driven LS-type PCA Algorithm results in suboptimal solution of (1) and (2). However, we may expect that Data-driven LS-type PCA's solution appear closely to that of classical PCA, since the data points tends to be clustered along principal directions. As partial verification, we executed a small Monte Carlo study designed as follows.

Step 1: Generate $n$ independent observations with $p$ variables $Z_1, \cdots, Z_p$ such that

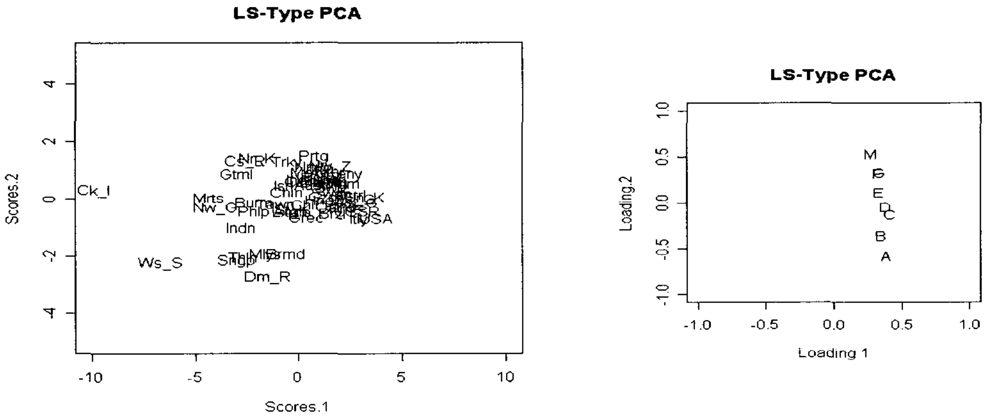
$$Z_j \sim N(0, j^2)\ , \quad j = 1, \cdots, p$$

to form the $n \times p$ data matrix $Z$.

Step 2: Without any prior adjustment, calculate the first and the second eigenvectors $v_1$ and $v_2$, of $Z^t Z/n$. Also, obtain $v_1^*$ and $v_2^*$ by Data-driven LS-type PCA.



<Figure 1> Classical PCA plots of National Track Records Data



<Figure 2> Data-driven LS-type PCA plots of National Track Records Data

Step 3: Compute

$$\text{cosines.1} = |\, v_1^t v_1^* \,| \quad \text{and} \quad \text{cosines.2} = |\, v_2^t v_2^* \,|$$

for closeness measures between two loading vectors of PC dimensions 1 and 2. The measures range from 0 (=far from close) to 1 (=completely close).

In Experiment 1, we set $p=5$ and $n=400/100/40$. The 1,000 simulation results are summarized in <Table 1>. For the case $n=400$, the result is perfectly favorable to our claim. But we should expect that the result worsens as $n$ decreases. Even for the case $n=40$, however, the result seems to be fine for practical use.

In Experiment 2, we set $p=10$ and $n=400/100/40$. The 1,000 simulation results are summarized in <Table 2>. The case in which $p=10$ and $n=40$ is the worst. The reason is obvious: we have only four observations per dimension $(n/p=4)$.

<Table 1> Simulation result for the case $p$ =5.

| | cosines.1 | | cosines.2 | |
|---|---|---|---|---|
| $n$ = 400 | average 0.992 | sd 0.005 | average 0.992 | sd 0.006 |
| $n$ = 100 | average 0.983 | sd 0.012 | average 0.982 | sd 0.014 |
| $n$ = 40 | average 0.971 | sd 0.030 | average 0.965 | sd 0.035 |

<Table 2> Simulation result for the case $p$ =10.

| | cosines.1 | | cosines.2 | |
|---|---|---|---|---|
| $n$ = 400 | average 0.939 | sd 0.054 | average 0.922 | sd 0.065 |
| $n$ = 100 | average 0.911 | sd 0.068 | average 0.875 | sd 0.104 |
| $n$ = 40 | average 0.897 | sd 0.086 | average 0.837 | sd 0.132 |

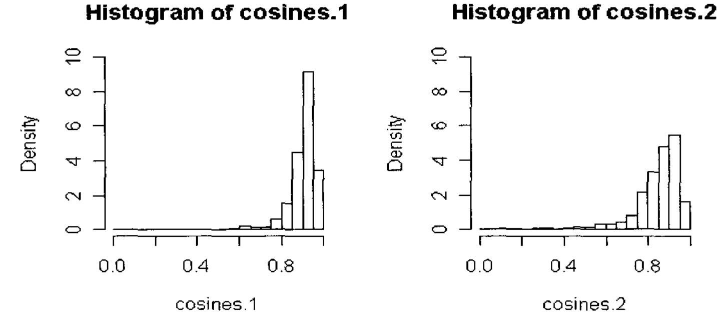
Even in that case, cosines.1 is close to 0.9 on average. <Figure 3> shows the distribution of simulation outcomes for the worst case ($p=10$ and $n=40$).

In sum, we may conclude that the data-driven optimization works well, unless $n/p$ is too small (say less than 5).

Our PCA algorithm proposed above is a modification of Hubert et al. (2002) and Croux and Ruiz-Gazen (2005), of which the LS version could be written as the maximization of the dispersion measures of principal component scores:

$$\sum_{i=1}^{n} \| (x_k^t x_i)/(x_k^t x_k)\, x_k \|^2, \quad \text{for } k=1,\cdots,n. \tag{4}$$

Certainly, the minimization of (3) and the maximization of (4) are equivalent each other when the sum of squares or the "mean" squares criterion is adopted, but they are not so if the mean is replaced by median or if the sum is replaced by trimmed sum. That motivates LMS and LTS-type versions, which we will develop in the next section.



<Figure 3> Simulation outcomes for the case $p$ =10 and $n$ =40.

## 3. Data-driven LMS and LTS-type PCA's

We propose "Data-driven Least Median of Squares (LMS)-type" PCA algorithm, that is same as Data-driven LS-type PCA with one exception: Replace (3) by

$$\text{median}\{ \parallel x_i - (x_k^t x_i)/(x_k^t x_k)\, x_k \parallel^2;\ k = 1,\cdots,n \}. \tag{5}$$

Similarly, we propose "Data-driven Least Trimmed Squares (LTS)-type" PCA algorithm, which is same as Data-driven LS-type PCA with one exception: Replace (3) by

$$\sum_{i=1}^{n-h} \parallel x_i - (x_k^t x_i)/(x_k^t x_k)\, x_k \parallel^2 \tag{6}$$

where $h$ is an integer between 0 and $[n/2]$, where $[m]$ is the largest integer less than or equal to $m$. In this study, $h$ is set equal to $[n/4]$ or 25% of the sample size, for balancing maximum breakdown level (50%) and full Gaussian efficiency in the case of no contamination (0%). As in linear regression, we may expect that Data-driven LMS-type PCA has the maximum breakdown property but lacks efficiency and that Data-driven LTS-type PCA possesses the balance between robustness and efficiency.

Our criteria (5) and (6) differ from that of Hubert et al. (2002) and Croux and Ruiz-Gazen (2005). Their criterion is the maximization of robust dispersion measure such as

$$Q_n(s_1,\cdots,s_n) = 2.2219 \quad c_n \{\ |\ s_i - s_j\ |\ ; i < j \}_{(k)}, \tag{7}$$

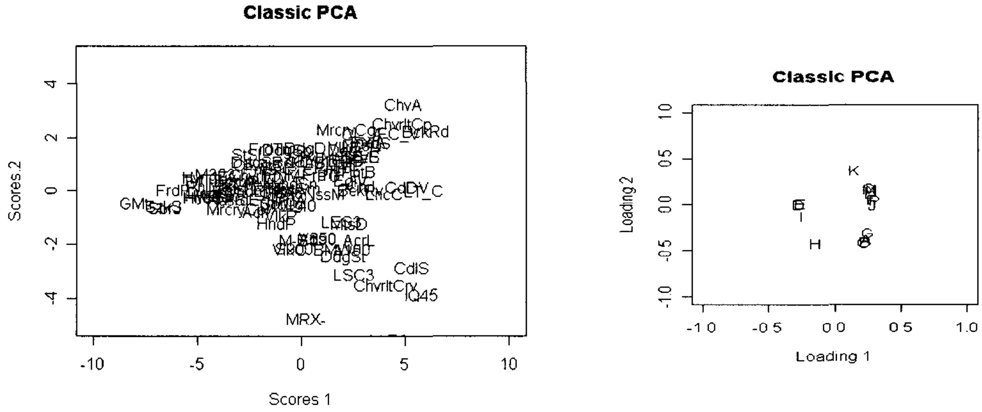
where $s_1 = v^t x_1, \cdots, s_n = v^t x_n$ are principal components scores, $c_n$ is a small-sample correction factor converging to 1 as $n$ increases,

$$k = \binom{h}{2} \quad \text{and} \quad h = \left[\frac{n}{2}\right] + 1 .$$

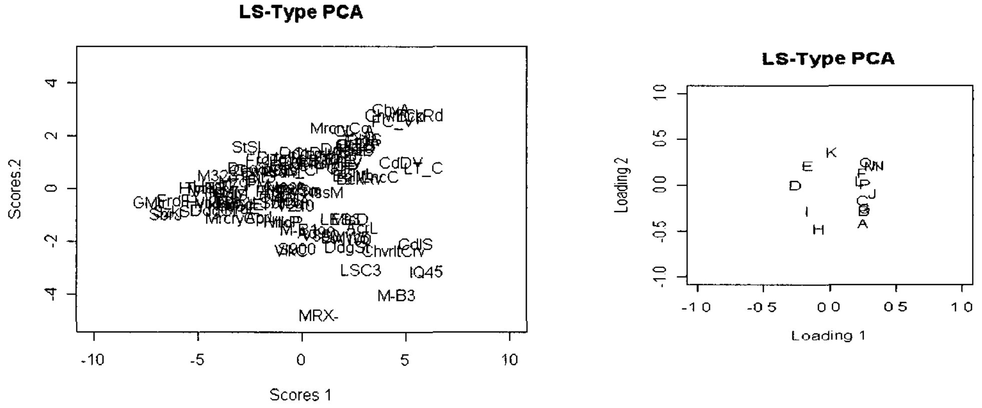
We think that minimization of (5) and/or (6) is more appropriate, instead of maximization of (7), since PCA can be viewed primarily as a dimensional reduction technique.

Applied to the inversely transformed National Track Records data of Section 2, Data-driven LMS and LTS-type PCA yield the same plot as that of Data-driven LS-type PCA, that is <Figure 2>. Thus, for the National Track Records data, different criteria of optimization (LS/LMS/LTS) do not affect the result. But this is a rather exceptional case.

We applied Classical, Data-driven LS-type, LMS-type and LTS-type PCA to Cars93 dataset which consists of 93 observations (=automobile makers/models) with 16 variables for various automobile attributes such as min-price "A", price "B", max-price "C", mpg-city "D", mpg-highway "E", engine size "F", horsepower "G", rpm "H", rev-per-mile "I", fuel-tank-capacity "J", passengers "K", length "L", wheel

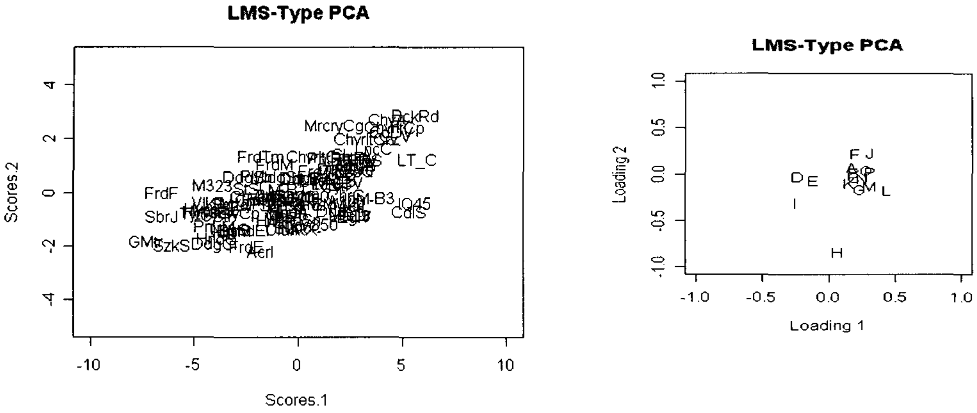


<Figure 4> Classical PCA plots of Cars93 data

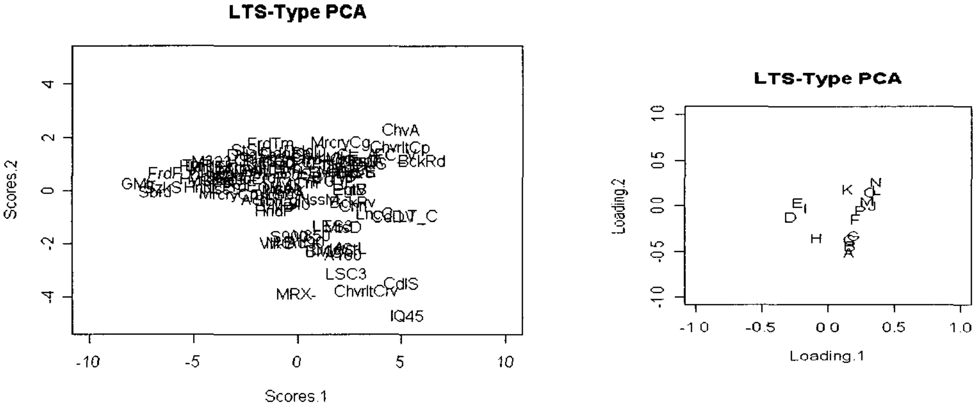


<Figure 5> Data-driven LS-type PCA plots of Cars93 data

base "M", width "N", turn-circle "O", and weight "P". The results are shown in <Figures 4, 5, 6, 7>. One may see that the pictures in <Figures 4, 5, 7> are similar: 1) Economy small cars position in the left edge in scores plot, 2) large inexpensive cars in the upper right corner, and 3) high-function expensive cars in the lower right corner. In <Figure 6> of Data-driven LMS-type PCA, the automobile points runs in curved band from the left end with economy small cars to the upper right end with large expensive cars. The dataset and full names of automobile makers/models are available from MASS library of R software (http://www.r-project.org).



<Figure 6> Data-driven LMS-type PCA plots of Cars93 data



<Figure 7> Data-driven LTS-type PCA plots of Cars93 data

## 4. Concluding Remark

The techniques of this study for PCA may be classified into projection-pursuit methodology (Li and Chen, 1985). Alternative approach for robust PCA is based on robust covariance estimation procedures such as Minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) (Rousseeuw 1984, Rousseeuw and Leroy 1987). Apparent difference between two methods is that the former is directly focused on the reduction of dimensions into two, three or so, while the latter is broadly targeted to full dimensional covariance property. For the case $n < p$, only the former method works.

In the case of data sets with large sample size $n$ , computational burden required for producing a data-driven PCA is quite moderate since the proposed algorithm contains finite number of non-nested iterative loops of length $n$ . For data sets with extremely large $n$ , one may further reduce the computational burden by sampling subset of observations to derive principal axes.

## References

[1] Croux, C., and Ruiz-Gazen, A. (1996). A fast algorithm for robust principal components based on projection pursuit. *COMPSTAT Proceedings in Computational Statistics*. Physica-Verlag, Heidelberg. 211-216.

[2] Croux, C., and Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: The projection-pursuit approach revisited. *Journal of Multivariate Analysis*, Vol. 95, 206-266.

[3] Hubert, M., Rousseeuw, P.J., and Verboven, S. (2002). A fast method for robust principal components with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems*, Vol. 60, 101-111.

[4] Huh, M.H. (1999). *Quantification Methods for Multivariate Data.* Freedom Academy, Seoul. (Written in Korean)

[5] Johnson, R.A. and Wichern, D.W. (1992) *Applied Multivariate Statistical Analysis* (Third Edition). Prentice Hall, Englewood Cliffs.

[6] Lebart, L., Morineau, A., and Warwick, K. (1984). *Multivariate Descriptive Statistical Analysis.* Wiley, New York.

[7] 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,* Vol. 80, 759-766.

[8] Rousseeuw, P.J. (1984). Least median of squares regression. *Journal of the American Statistical Association,* Vol. 79, 871-880.

[9] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection. Wiley,* New York.

[Received March 2006, Accepted April 2006]