Cilt: 06 Sayı: 01 Kategori: 01 Sayfa: 18-37 ISSN No: 1303-6319 Volume: 06 Sayı: 01 Category: 01 Page: 18-37

THE PERFORMANCE EVALUATION OF ROBUST PAIRWISE COVARIANCE ESTIMATOR

Özlem YORULMAZ[*]

ABSTRACT

Multivariate analysis and multidimensional outlier detection techniques necessitate using robust high breakdown covariance estimators, which have time saving algorithms in the presence of outliers in high dimensional data. The preference for robust estimators arises from the distortion effect of outliers when classical estimators are used. Orthogonalized Gnanadesikan-Kettering (OGK) estimator (Maronna and Zamar, 2002) was devised in order to address the computational challenge of high breakdown estimators. In this study the focus is on the evaluation of some covariance estimators in Principal Component Analysis (PCA). A comparison of the performance of OGK in PCA and Robust Principal Component Analysis (ROBPCA) (Hubert et al, 2005) has been carried out by way of simulations and with real data sets.

Key Words: Fast minimum covariance determinant estimator, Orthogonalized Gnanadesikan-Kettering estimator, Outliers, Principal components analysis, Robust principal component analysis.

1. INTRODUCTION

As is commonly known, the covariance matrix is one of the fundamental instruments of statistical analysis that is widely used for obtaining correlation coefficients between variables, reducing the number of variables and diagnosing multivariate outliers.

An observation whose pattern differs from the majority of data is generally called an outlier. Outliers may cause misleading estimations when classical empirical covariance matrices are used; therefore, statisticians directed their attention to robust techniques and different robust methods have been invented to estimate the covariance matrix. If there are some outliers in the data, the classical (maximum likelihood) estimator of the covariance matrix may not prevent masking (case when analysis suggests that one or more outliers are in fact good cases) and swamping (case when analysis suggests that one or more good cases are outliers) effects. For this reason it is much safer to use robust estimators instead.

[*] Istanbul University, Faculty of Economics, Department of Econometrics, e-mail: yorulmaz@istanbul.edu.tr

Here some estimators are defined briefly; throughout the definitions X_{nxp} notation is used which stands for nxp data matrices, where n indicates the number of objects and p indicates the number of variables.

OGK is a robust covariance matrix estimator for high dimensional data sets which has been proposed by Maronna and Zamar (2002) as an alternative to Fast Minimum Covariance Determinant (FMCD) estimator. FMCD (Rousseeuw and Van Driessen, 1999) is a high breakdown robust estimator, an improved form of the Minimum Covariance Determinant (MCD) high breakdown estimator. It has been stressed by Maronna and Zamar (2002) that the increase of cases (n) diminishes the high breakdown property of FMCD and it also has been emphasized that the increase of the dimension (p) requires immense computational time for FMCD although it is the quicker alternative of MCD.

Underlying purposes of the study are firstly using OGK covariance matrix in one of the dimension reduction technique PCA, secondly comparing and evaluating some properties of robust and classical matrices with several data sets and simulations. Through the comparison and evaluation step a matlab Library for Robust Analysis (LIBRA) was used and besides codes for OGK estimator were written in Matlab (See, Appendix-2).

2. PROPERTIES OF ROBUST ESTIMATORS

The properties of robust covariance estimator can be summarized as breakdown value, positive definiteness and affine equivariance. These properties allow a characterization of estimators as low breakdown, high breakdown, affine and not affine.

Breakdown value is a maximum amount of contamination that an estimator can carry. This value also measures the robustness of an estimator. As can be inferred from the following notations,

$\hat{\Sigma}$: Covariance matrix estimator
X : Data matrix,
X': Matrix obtained by replacing m points out of X

the breakdown value, ε_n^*, is the largest eigenvalue of $\hat{\Sigma}$ driven to ∞ or the smallest eigenvalue of $\hat{\Sigma}$ driven to zero:

$$\varepsilon_n^*(\hat{\Sigma}, X) = \min\left\{\frac{m}{n}\sup_{x'}\frac{\lambda_{\max(X')}}{\lambda_{\min(X')}} = \infty\right\}$$

Conventional wisdom tells that the covariance matrix yields multivariate scatter of data which is represented by an ellipsoid. The affine equivariance and positive definiteness properties that were mentioned above are strongly related to this ellipsoid because the eigenvectors of a covariance matrix determine the axes of an ellipsoid and the eigenvalues of this covariance matrix are equal to the length of these axes. Given this geometrical concept, the positive definiteness of a covariance matrix can be easily perceived.

Generally the location and scale estimators are expected to be affine equivariant, which means that after a linear transformation of the data the estimators will be transformed accordingly. If A_{pxp} is an orthogonal matrix $(A' = A^{-1})$ and the data matrix is transformed as $XA' + 1_n v$, then the center $\hat{\mu}_x$ and the loading matrix $P_{p,k}$ of CPCA or ROBPCA are equal to $A\hat{\mu}_x + v$ and AP respectively. The eigenvalues of the defined ellipsoid and the scores remain the same under this transformation for CPCA and ROBPCA. If an orthogonal transformation is applied to the data as XA' and an estimator rotates accordingly, this estimator can be defined as an orthogonal equivariant. From the above it can be deduced that CPCA and ROBPCA estimators are location and orthogonal equivariant but, as will become clear from the simulation study, OGK is not. This can be rated as a disadvantage of OGK because the absence of the equivariance property makes it hard to predict the behavior of the OGK against outliers on rotated data.

3. ADVANCES ON ROBUST COVARIANCE ESTIMATORS

In the statistical literature, a substantial number of studies have been proposed about robust scatter matrix estimation. The M estimator is the initial robust scatter matrix estimator which was suggested by Hampel in 1973, then studied by Maronna (1976) and Huber (1981). This estimator is positive definite and affine equivariant, but its breakdown point, 1/p, is not satisfactory.

Subsequently, high breakdown affine equivariant and positive definite estimators have been studied. These are the Stahel-Donoho (SD) estimator by Stahel-Donoho (1981) and studied by Maronna and Yohai (1995), the Minimum Volume Estimator (MVE) and the Minimum Covariance Determinant (MCD) by Rousseeuw (1987, 254). Due to the efficiency in high dimensions Croux and Haesbroek recommend to use MCD instead of MVE (2000).

MCD is a highly robust estimator of multivariate location and scatter. Its objective is to find h observations out of n whose classical covariance matrix has the lowest determinant where h is defined as a default value (n+p+1)/2. The value for h is $[n+p]/2 \leq h \leq n$. The estimation of MCD is time-consuming and therefore limited to a few hundred objects in a few dimensions since the exact solution has to be found among all possible subsets of n observations taken in h dimensional subgroups.

FMCD (Rousseeuw and Van Driesen, 1999) has been developed to address this shortcoming; the algorithm of this estimator is set up on a re-sampling scheme which is called the C-step. But it has to be stressed that FMCD still requires substantial computation time when n is large (Alqallaf et al, 2002).

4. OGK ESTIMATOR

As a result of giving up the requirements of affine equivariance and positive definiteness, one can get estimates much faster. A straightforward estimator for multivariate location is a coordinatewise one which can be calculated from a robust location estimator for each variable in the data. Similarly for a multivariate covariance matrix, pairwise estimators can be used by applying a robust covariance estimate to each pair of variables.

Because of the computational burden of affine equivariant and high breakdown estimators, Marrona and Zamar (2002) dropped the affine equivariance property and introduced the OGK estimator. OGK is based on the Gnanadesikan-Kettenring (G-K) robust pairwise covariance matrix estimate. The G-K estimator,

$$\operatorname{cov}(X,Y)=\frac{1}{4}(\sigma(X+Y)^2-\sigma(X-Y)^2)\,,$$

was suggested by Gnanadesikan and Kettenring (1972). This estimator is not guaranteed to be positive definite whereas the OGK pairwise estimator preserves positive definiteness.

Before explaining the steps of the algorithm, some notations have to be defined as X_j refers to the columns of X data matrix where j is $j=1,.....p$ and $x_i^{'}$ refers to the rows where i is $i=1,...n$.

- For each variable MAD values and w_i weights are calculated. MAD stands for the median absolute deviation from the median and w_i values are obtained from $W_c(x)$ function.

  $$\sigma_{0j}=MAD(X_j)=med(\left|X_j-1_n med(X_j)\right|)$$

  $$W_j=W_{c1}\left((x_j-1_n med(X_j))/\sigma_{0j}\right),\ W_c(x)=\left(1-\left(x/c\right)^2\right)^2 I(\left|x\right|\le c) \qquad (1)$$

  and *I (.)* is the indicator function.

- Location and scale statistics are obtained from

  $$\mu(X_j)=\sum_i x_{ij}w_{ij}/\sum_i w_{ij} \qquad \text{and}$$

  $$\sigma(X_j)^2=(\sigma_{0j}^2/n)\sum_i \rho_{c_2}\left((x_{ij}-\mu(X_j))/\sigma_{0j}^2\right) \qquad (2)$$

  where ρ_{c_2} can be obtained from $\rho_c(x)=\min(x^2,c^2)$.

Maronna and Zamar (2002) proposed to use $c_1=4.5$ and $c_2=3$ for combining the robustness and efficiency.

- A new diagonal matrix is defined by means of scale statistics that were obtained in the previous step

  D = diag($\sigma(X_1),....,\sigma(X_p)$); using the inverse of D with the columns of X_j a new variable,

$$Y = D^{-1}X' , \qquad (3)$$

Y is defined.

This step makes the estimator scale equivariant.

- U=[u_{jk}] correlation matrix is computed by applying v to the columns of Y.

$$U_{jk} = v(Y_j, Y_k) = \begin{cases} \frac{1}{4}(\sigma(X_j + Y_k)^2 - \sigma(X_j - Y_k)^2) & j \neq k \\ 1 & j = k \end{cases} \qquad (4)$$

- The eigenvalues λ_j and the eigenvectors e_j of U ($j = 1,....p$) are obtained and new matrices are defined as $\Lambda = diag(\lambda_1,..,\lambda_p)$ and E whose columns are the e_j 's. Then U is decomposed as $U = E\Lambda E'$.
  $A = DE$, and $Z = (E'Y')' = (A^{-1}X')'$ are defined. (5)

- After the extraction of $\Gamma = diag(\sigma(Z_1)^2, ...,\sigma(Z_p)^2)$, the seeking Orthogonolized Gnanadesikan-Kettenring estimators $V(X) = A\Gamma A'$ and $t(X) = Av$, where and
  $v = (\mu(Z_1),....,\mu(Z_p))'$ are found. (6)

- Maronna and Zamar (2002) suggested using an improvement for the resulting estimator by a reweighting procedure.

$$t_{wj} = \sum_i w_i X_j / \sum_i w_i \ , \qquad V_{wj} = [\sum_i w_i (X_j - 1_n t_{wj})(X - 1_n t_{wj})'] / \sum_i w_i \quad (7)$$

The weight function W, $W(d) = I(d \leq d_0)$, can be extracted from,

$$d_i = \sum_j \left((z_{ij} - \mu(Z_j)) / \sigma(Z_j) \right)^2 \text{ and } \quad d_0 = \chi_p^2(\beta) med(d_1,..,d_n) / \chi_p^2(.5)$$

This resulting estimator is called R-OGK (Reweighted Orthogonolized Gnanadesikan-Kettenring) estimator.

Maronna and Zamar discussed different β values with respect to their simulation results and they mentioned that β *=0.90* generally yielded the best results. Also the R-OGK procedure can be iterated by replacing U in step 5 by $E\Gamma E'$ until convergence but authors warned not to iterate beyond the second iteration.

5. CLASSICAL AND ROBUST PCA

Principal Component Analysis is a technique for explaining the covariance structure of the data by forming new orthogonal variables which are linear combinations of the original variables. These new variables are referred to as principal components which correspond to the eigenvectors of the covariance matrix. The first principal component accounts for the maximum variance of projected data points on it. The second principal component accounts for the maximum variance that has not been accounted for by the first principal component. The procedure continues in this way and it is expected to use few principal components for most of the variance in the data.

But as the principal components are the eigenvectors of classical covariance matrix, it is possible that the components have been adversely influenced by outliers. In this case it is preferable to use robust principal component approaches which can prevent outlier effects. These approaches can be categorized into three different groups:

- replacing classical covariance matrix with robust covariance matrix
- using projection pursuit method
- combining projecting pursuit and robust covariance matrix

Campbell (1980) used M estimators of covariance matrices but they are not resistant against many outliers. Croux and Haesbroeck (2000) used MCD by replacing the classical covariance matrix. However this method is limited to small, moderate samples. In this study, the OGK covariance matrix was replaced with the classical covariance matrix in a similar way and the results are presented through simulation and real data sets.

Li and Chen (1985) and Hubert et all (2002) used the projection pursuit method for obtaining robust PCA.

Hubert, Rousseeuw and Vanden Branden (2005) proposed ROBPCA method which is a combination of the projection pursuit method and the MCD estimator.

6. EVALUATION OF THE ESTIMATORS’ PERFORMANCE

The assessment of breakdown point and computational time of CPCA, ROBPCA, PCA with OGK and R-OGK were carried out on real data sets and with simulations.

Before illustrating the methods on real data sets, it is necessary to mention the kind of outliers that can occur and their diagnostic plot. Here the definitions are given briefly. A satisfactory explanation with a visual plot can be found in ROBPCA (Hubert et al, 2005).

- Good leverage points: These points lie close to the PCA space but far from the major homogenous data group.
- Orthogonal outliers: These observations have large orthogonal distances to the PCA space; only their projections can be seen on the PCA space.

- Bad leverage points: This type of observations has a large orthogonal distance and its projections on the PCA space are far from typical projections.

The classification of observations can be identified from a diagnostic plot. The horizontal axis of the diagnostic plot consists of the score distance and the vertical axis of the diagnostic plot consists of the orthogonal distance.

- Score distance is calculated for each observation with

$$SD_i = \sqrt{\sum_{j=1}^{k} (t_{ij}^2 / l_j)}$$

where the t_{ij} pca scores are obtained from $T_{n,k} = (X_{n,p} - 1_n \hat{\mu}')P_{p,k}$. Here, $l_1,...l_k$ stands for the eigenvalues and $P_{p,k}$ represents the matrix which consists of eigenvectors.

- Orthogonal distance is defined for each observation as

$$OD_i = \left\| x_i - \hat{\mu} - P_{p,k} t_i' \right\|$$

For classifying observations two cut-off lines are drawn. The cut off value on the horizontal axis is $\sqrt{\chi^2_{k,0.975}}$. There are several approaches for the distribution of the cut-off value on the vertical axis (Hubert et al, 2005). According to the Wilson-Hilferty approach orthogonal distances to the power 2/3 are normally distributed. Estimations of the mean and the variance of this distribution were found by means of univariate MCD in ROBPCA paper, in a similar way the $\hat{\mu}$ and $\hat{\sigma}^2$ for OGK and R-OGK were found by univariate OGK and univariate R-OGK. Then, the cut-off value on the vertical axis is defined as $(\hat{\mu} + \hat{\sigma} z_{0.975})^{3/2}$.

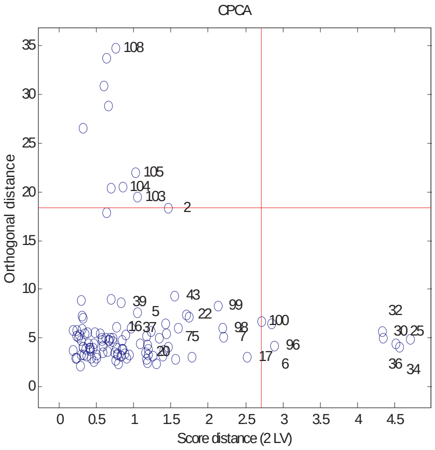
6.1 Real Data

CPCA, ROBPCA, OGK and R-OGK methods were applied on two data sets[*] which are commonly used in robust studies.

6.1.1 Car data

The first example is the low dimensional car data set which contains 111 cars and 11 different characteristics of cars. From the Figure 1 observations 25, 30, 32, 34, 36 are seen as good leverage points and observations 103-108, 110 are seen as orthogonal outliers. However the diagnostic plots of ROBPCA (Figure 2) and R-OGK (Figure 3) identifies this orthogonal outlier group and observations 106, 108 and 110 as bad leverage points. OGK (Figure 4) also converts those cases to bad leverage points but with a difference. As it seen from the Figure 2 they are very close to boundary.

[*] Data were provided by Karlien Vanden Branden



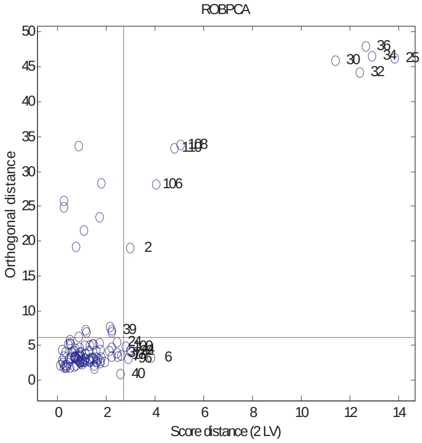
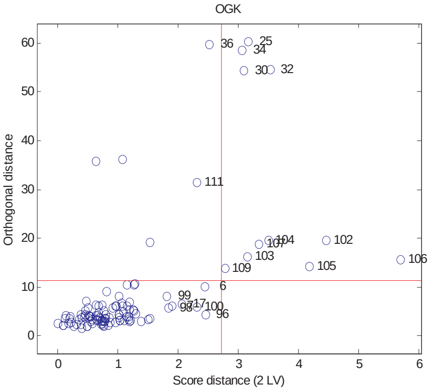


Figure 1. Diagnostic plot of car data set based on two CPCA Principal Components

Figure 2. Diagnostic plot of car data set based on two ROBPCA Principal Components



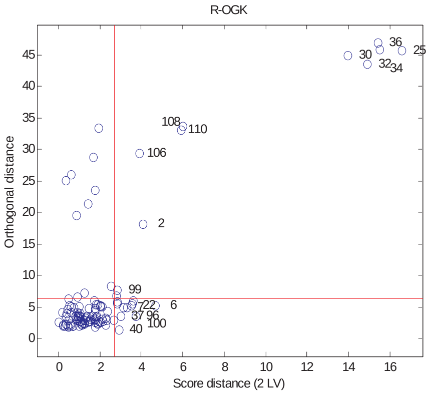


Figure 3. Diagnostic plot of car data set based on two OGK PCA Principal Components

Figure 4. Diagnostic plot of car data set based on two R-OGK PCA Principal Components

6.1.2 Octane data

The second example is the Octane high dimensional data set which consists of 226 variables and 39 gasoline samples. In this data set, six samples contain (25, 26, 36-39) added alcohol.

It is obvious from Figure 5 that CPCA can detect only outlying 26 as a bad leverage point. In contrast, ROBPCA (Figure 6), OGK (Figure 7) and ROBPCA (Figure 8) find all outlying points. This shows that ROBPCA, OGK and R-OGK methods do not contain outliers in their estimated subspaces.

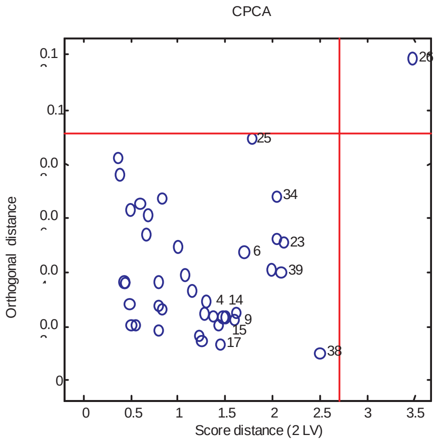


Figure 5. Diagnostic plot of octane data set based on two CPCA Principal Components

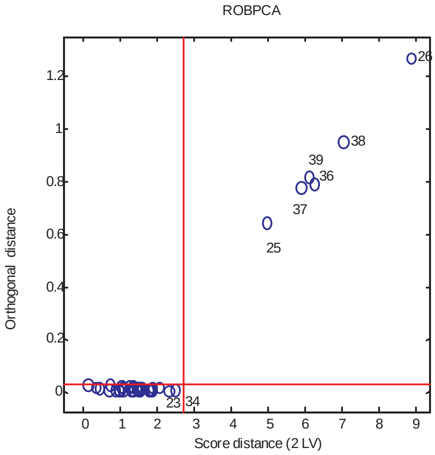


Figure 6. Diagnostic plot of octane data set based on two ROBPCA Principal Components

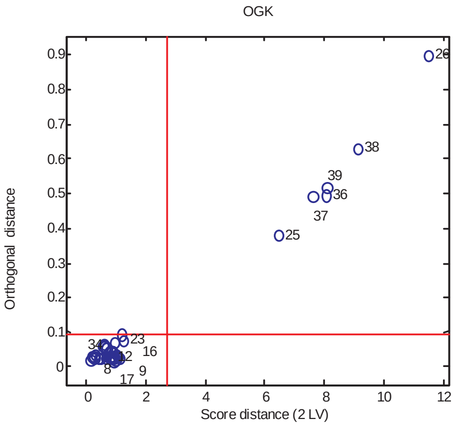


Figure 7. Diagnostic plot of octane data set based on two OGK PCA Principal Components

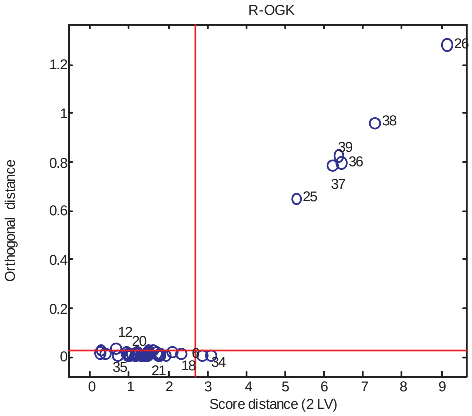


Figure 8. Diagnostic plot of octane data set based on two R-OGK PCA Principal Components

Contrary to the car data, this time OGK on high dimensional data showed similarity to ROBPCA and R-OGK.

6.2 Simulation

In this section a simulation study is performed to compare the performances of CPCA, ROBPCA, OGK and R-OGK on low and high dimensional data sets.

While generating the data, the following contaminated model construction is used

$$(1-\varepsilon)N_p(0,\Sigma)+\varepsilon N_p(\tilde{\mu},\tilde{\Sigma})$$

with different values for epsilon and different sizes of the data matrix. $\tilde{\mu}$ represents the center of outliers and is adjusted to obtain bad leverage points as will become apparent in the following.

For each setting 100 data sets were constructed and two different assessment criteria, MAXSUB and MSE, were used to gain insights about their performance. MAXSUB is the maximal angle between the space spanned by the estimated principal components and E_k, where E_k is the subspace spanned by the *k* dominant eigenvectors of Σ.

The MAXSUB measure is defined as (Hubert et al, 2005) $I'_{k,p}P_{p,k}P'_{k,p}I_{p,k}$ $\text{MAXSUB}=\arccos(\sqrt{\lambda_k})$ where λ_k is the smallest eigenvalue of $I'_{k,p}P_{p,k}P'_{k,p}I_{p,k}$. This gives the largest angle between a vector in E_k and the vector most parallel to it in the estimated PCA subspace. MAXSUB provides the best values when it is close to 0.

The second criterion, MSE, is the mean squared error of *k* largest eigenvalues and defined as:

$$\text{MSE}(\hat{\lambda}_j)=\frac{1}{100}\sum_{l=1}^{100}(\hat{\lambda}_j^{(l)}-\lambda_j)^2$$

Due to their lacking the orthogonal equivariance property, the performance of OGK and R-OGK estimators has also been evaluated on a rotated data matrix which has been obtained by multiplying the original data matrix with an orthogonal matrix.

6.2.1 Simulation study when $\varepsilon=0.20$ and $\varepsilon=0.10$ in low dimension

These are the assigned values of parameters that used for generating low dimensional settings:
n=150, p=5, $\Sigma=\text{diag}(12,8,6,0.20,0.05)$, k=3. It has been decided to assign a value of 3 to k, because three components explain 99% of the data ($(\sum_{i=1}^{3}\lambda_i)/(\sum_{i=1}^{5}\lambda_i)=0.9905$).

As can be seen from Figure 9 and Figure 10 the worst MAXSUB value pertains to CPCA; it is close to 1 when 20% contamination is added to the data. ROBPCA gives the best result and R-OGK pursuits ROBPCA. The most striking result here is that R-OGK is much more equivariant than OGK after rotation, the estimations of R-OGK on rotated data matrix are approximately equivariant. In case of a 10% contamination of the data, OGK is very much in line with the other estimators. Tables 3, 4 give exact values of MAXSUB.

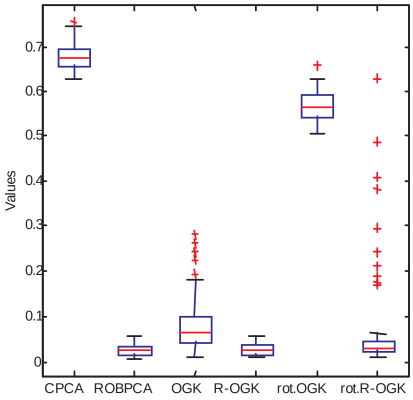


Figure 9. Boxplots of 20% contaminated low dimensional dataset based on MAXSUB values

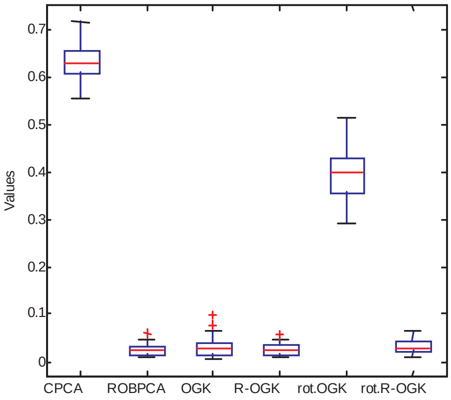


Figure 10. Boxplots of 10% contaminated low dimensional dataset based on MAXSUB values

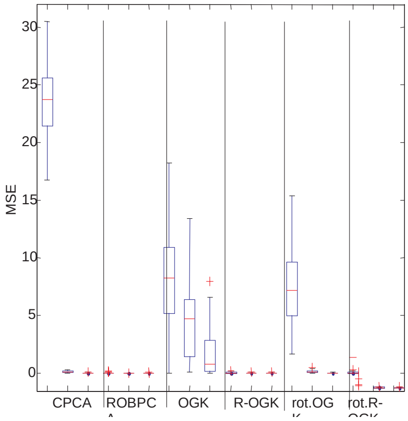


Figure 11. Boxplot of 20% contaminated low dimensional data set based on MSE of eigenvalues

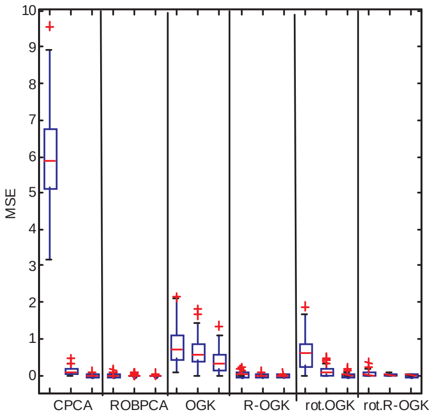


Figure 12. Boxplot of 10% contaminated low dimensional data set based on MSE of eigenvalues

Figures 11 and 12 just enable to evaluate the first eigenvalues but Tables 7, 8 provide detailed information for the MSE of three eigenvalues from which it becomes evident that ROBCPA gives the best results and R-OGK is next in ranking.

6.2.2 Simulation study when $\varepsilon = 0.20$ and $\varepsilon = 0.10$ in high dimension

In high dimensional simulation studies the following parameter values were used: p=100, n=50, $\tilde{\mu} = (6,8,10,12,14,16,0,0....0)$, $\Sigma = (12,8,6,5,3,0.1,0.099,0.098.....0.006)$ and k=5. The first five eigenvalues explain 87% of the data $((\sum_{i=1}^{5}\lambda_i)/(\sum_{i=1}^{100}\lambda_i) = 0.8710)$.

For high dimensional data, MAXSUB values of OGK give surprising results which are evident from Figure 13 and Figure 14. Contrary to what is deduced from the MAXSUB values, the MSE of the eigenvalues indicates that OGK fails like CPCA. R-OGK and ROBPCA, however, give similar and best results for both criteria (Figure15, 16).

Based on the MAXSUB values, OGK on rotated data matrix breaks down. This is in contrast to the MSE of the eigenvalues which tells that the worst outcome is pertained with OGK (See, Appendix Table 5, 6 and Table 9, 10). So there is a serious contradiction between the MAXSUB and MSE values for OGK. When the results of two criteria (MAXSUB and MSE of eigenvalues) are compared, it has to be noticed that except OGK and OGK on rotated X, all the other estimators give coherent results with each other.

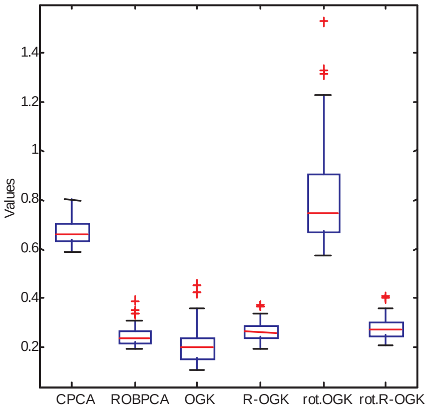


Figure 13. Boxplots of 20% contaminated high dimensional dataset based on MAXSUB values

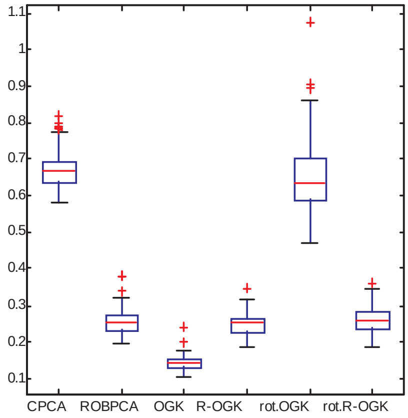


Figure 14. Boxplots of 10% contaminated high dimensional dataset based on MAXSUB values

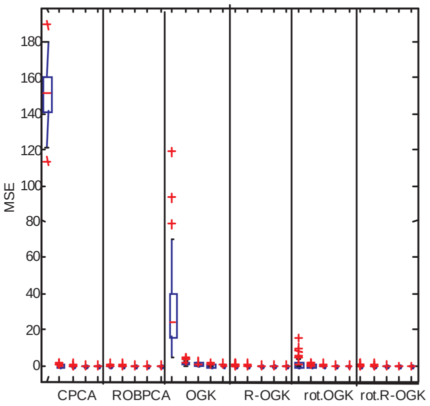


Figure 15. Boxplot of 20% contaminated high dimensional data set based on MSE of eigenvalues

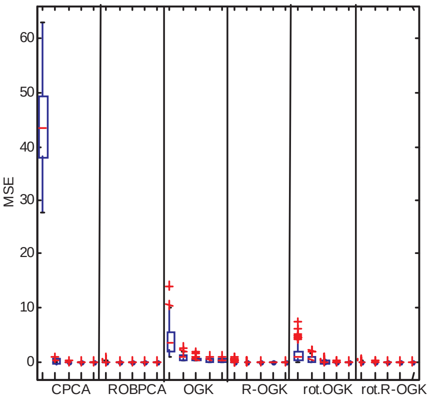


Figure 16. Boxplot of 10% contaminated high dimensional data set based on MSE of eigenvalues

6.2.3 Simulation study when $\varepsilon = 0$ in high and low dimension

For uncontaminated data in the high and low dimensional case CPCA and OGK give the best results, with OGK even performing slightly better than CPCA. Although the MAXSUB results show very similar performances with respect to those of the ROBPCA and the R-OGK estimators, the MSE results indicate that ROBPCA is better. ROBPCA and R-OGK yield higher MAXSUB and MSE values in comparison with lower dimension. OGK and R-OGK estimates on rotated data matrix do not perform extremely different from the original data matrix. The visual and numerical illustrations are provided in the tables (See Appendix-1, Table 1, 2) and below figure 17, 18, and 19, 20.

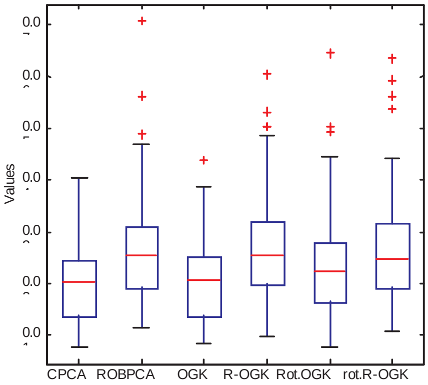


Figure 17. Boxplots of uncontaminated low dimensional dataset based on MAXSUB values

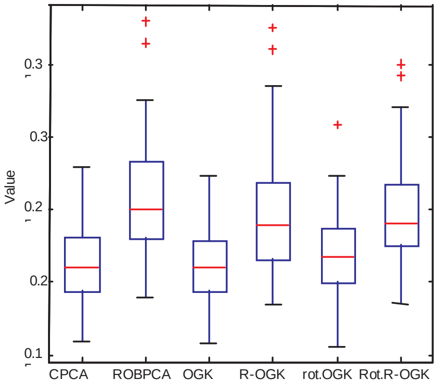


Figure 18. Boxplots of uncontaminated high dimensional dataset based on MAXSUB values

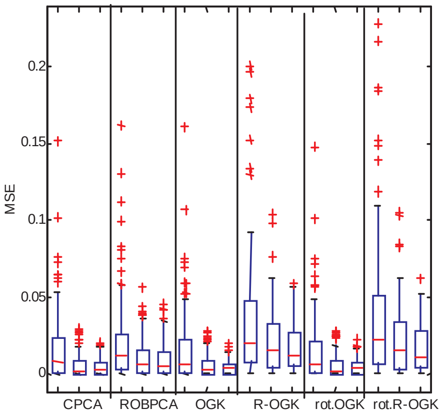


Figure 19. Boxplot of uncontaminated low dimensional data set based on MSE of eigenvalues

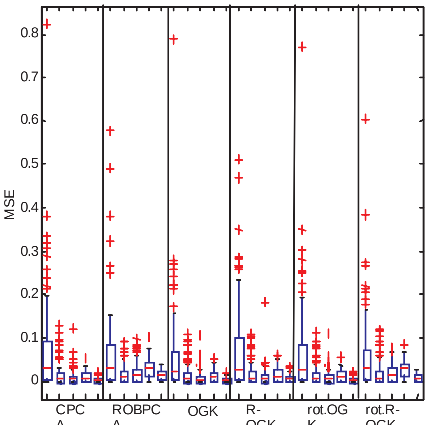


Figure 20. Boxplot of uncontaminated high dimensional data set based on MSE of eigenvalues

7. CONCLUSION

As a general result of simulation it can be said that, when there is contamination in the data, ROBPCA and R-OGK give very similar results, they are both superior to CPCA and OGK but in low dimension ROBPCA slightly comes into prominence whereas in high dimension R-OGK comes into prominence. So, when high dimension is the subject, it can be preferred to use R-OGK since it's computationally easier than ROBPCA. Furthermore, compared to OGK, R-OGK is more equivariant.

Nevertheless when there is no contamination in the data, CPCA and OGK yield best results. In this case inequivariance of OGK does not seem to be an important issue.

Another point, which should be stressed here, is that OGK shows the worst performance of robust estimators in contaminated data sets according to MSE criteria. But in contrast to MSE values, MAXSUB values specify the OGK estimator surprisingly as the best estimator especially in high dimensional data sets. The striking but inevitably incoherent differences between MSE and MAXSUB values of OGK can be seen in appendix–1.

8. REFERENCES

Alqallaf F.A, Konis K.P., Martin R.D. and Zamar R.H., 2002. Scalable robust covariance and correlation estimates for data mining. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 14-23.

Campbell N.A., 1980. Robust procedures in multivariate analysis I: Robust covariance estimation, Applied Statistics, 29, 3, 231-237

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, 603-618.

Gnanadesikan, R., and Kettenring, J.R., 1972. Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics, 28, 81-124.

Huber,P.J.,1981. Robust statistics, John Wiley&Sons, New York.

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, 60, 101–111.

Hubert M., Rousseeuw P. J., and Vanden Branden K., 2005. ROBPCA: A new approach to robust principal components analysis. Technometrics, 47:64–79.

Li, G., Chen, Z., 1985. Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo. J. Amer. Statist. Ass. 80, 759-766.

Maronna, R.A,.1976. Robust M-estimators of multivariate location and scatter, Ann. Stat., 4, 51-67.

Maronna, R.A. and Yohai, V. J., 1995. The Behavior of the Stahel-Donoho robust multivariate estimator. J. Amer. Statist. Assoc. 90, 330-341.

Maronna R.A. and Zamar R.H., 2002. Robust estimates of location and dispersion for high-dimensional data sets. Technometrics, 44, 307-314.

Rousseeuw P.J. and Leroy A. M., 1987. Robust regression and outlier detection. Wiley-Interscience, New York.

Rousseeuw P.J. and Van Driessen K.,1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41, 212–223.

Stahel W., 1981. Breakdown of covariance estimators, Research Report 31, ETH Zurich,fachgruppe fuer Statistik.

GÜÇLÜ İKİLİ KOVARYANS TAHMİNCİSİNİN PERFORMANS DEĞERLENDİRMESİ

ÖZET

Yüksek boyutlu veri kümelerinde aykırı gözlemlerin varlığı halinde, çok değişkenli analiz ve çok boyutlu aykırı gözlem teşhis teknikleri, zamanı etkin kullanan, kırılma noktası yüksek güçlü kovaryans tahmincilerin kullanımını zorunlu kılar. Klasik tahmincilerin aykırı gözlemler karşısında bozulması, güçlü tahmincilerin kullanımnıı gerektirir. FMCD kırılma noktası yüksek, yüksek boyutlu verilerde kullanımı uygun olan bir tahmincidir, fakat Maronna ve Zamar (2002), gözlem sayısının artmasıyla FMCD'nin önemli zaman aldığını ve yüksek kırılma noktasına sahip olma özelliğini yitirdiğini vurgular. OGK tahmincisi, yüksek kırılma noktasına sahip güçlü tahmincilerin işlem süresinin uzunluğu problemine yanıt vermek için (Maronna, Zamar, 2002) önerilmiştir. Bu çalışmada OGK tahmincisi ile çeşitli kovaryans tahmincilerinin performansı Temel Bileşenler Analizi (TBA) ile değerlendirilmiştir.

Anahtar Kelimeler: Aykırı gözlemler, Güçlü temel bileşenler analizi, Minimum kovaryans Determinat tahmincisi, Ortogonal Gnanadesikan-Kettering tahmincisi, Temel bileşenler analizi.

Appendix-1

Table 1. Simulation results of MAXSUB when there is no contamination in low dimension

	Mean	Median	Error
CPCA	0.021	0.0204	8.04E-04
ROBPCA	0.0266	0.0251	0.001
OGK	0.0209	0.0206	7.78E-04
R-OGK	0.0264	0.0253	9.42E-04
rot.OGK	0.0232	0.0222	9.55E-04
rot.R-OGK	0.0264	0.0246	9.85E-04

Table 2. Simulation results of MAXSUB when there is no contamination in high dimension

	Mean	Median	Error
CPCA	0.2136	0.2103	0.0024
ROBPCA	0.2569	0.2503	0.0035
OGK	0.2111	0.2095	0.0024
ROGK	0.2457	0.2390	0.0037
Rot.OGK	0.2190	0.2171	0.0026
rot.R-OGK	0.2480	0.2410	0.0034

Table 3. Simulation results of MAXSUB when there is 20% contamination in low dimension

	Mean	Median	Error
CPCA	0.6760	0.6732	0.0025
ROBPCA	0.0263	0.0241	0.0011
OGK	0.0809	0.0641	0.0057
R-OGK	0.0282	0.0257	0.0012
Rot.OGK	0.5669	0.5635	0.0031
Rot.R-OGK	0.0598	0.0310	0.0099

Table 4. Simulation results of MAXSUB when there is 10% contamination in low dimension

	Mean	Median	Error
CPCA	0.6348	0.6297	0.0033
ROBPCA	0.0259	0.0244	9.4153e-004
OGK	0.0300	0.0283	0.0016
R-OGK	0.0266	0.0255	0.0010
Rot.OGK	0.3977	0.4015	0.0051
Rot.R-OGK	0.0319	0.0297	0.0011

Table 5. Simulation results of MAXSUB when there is 20% contamination in high dimension

	Mean	Median	Error
CPCA	0.6736	0.6604	0.0049
ROBPCA	0.2471	0.2413	0.0034
OGK	0.2131	0.1983	0.0074
ROGK	0.2677	0.2638	0.0036
rot.OGK	0.8036	0.7473	0.0178
rot.R-OGK	0.2797	0.2701	0.0046

Table 6. Simulation results of MAXSUB when there is 10% contamination in high dimension

	Mean	Median	Error
CPCA	0.6719	0.6675	0.0050
ROBPCA	0.2567	0.2517	0.0035
OGK	0.1441	0.1437	0.0019
rOGK	0.2508	0.2514	0.0030
rot.OGK	0.6530	0.6356	0.0098
rot.R-OGK	0.2629	0.2575	0.0036

Table 7. Simulation results for MSE of eigenvalues when there is 20% contamination in low dimensional data set

	Mean			Median			Error		
CPCA	23.6913	0.1231	0.0208	23.7911	0.1120	0.0134	0.2771	0.0078	0.0022
ROBPCA	0.0219	0.0081	0.0059	0.0080	0.0046	0.0028	0.0033	0.0010	0.0008
OGK	8.1342	4.5081	1.593	8.2647	4.75	0.7506	0.397	0.311	0.1845
R-OGK	0.0416	0.0169	0.0139	0.0302	0.0096	0.0093	0.0042	0.0024	0.0014
rot.OGK	7.3213	0.1334	0.0181	7.1625	0.1200	0.0106	0.2825	0.0102	0.0018
rot.R-OGK	0.0708	0.0301	0.0163	0.0326	0.0210	0.0127	0.0160	0.0027	0.0014

Table 8. Simulation results for MSE of eigenvalues when there is 10% contamination in low dimensional data set

	Mean			Median			Error		
CPCA	5.9608	0.1181	0.0191	5.8622	0.0985	0.0120	0.1304	0.0082	0.0022
ROBPCA	0.0184	0.0102	0.0078	0.0089	0.0049	0.0034	0.0026	0.0013	0.0009
OGK	0.8031	0.6472	0.3513	0.6913	0.5975	0.3104	0.0482	0.0357	0.0268
R-OGK	0.0453	0.0198	0.0147	0.0279	0.0125	0.0083	0.0049	0.0021	0.0014
Rot.OGK	0.6555	0.1342	0.0400	0.6208	0.1129	0.0199	0.0429	0.0099	0.0042
Rot.R-OGK	0.0619	0.0399	0.0260	0.0383	0.0313	0.0229	0.0065	0.0033	0.0019

Table 9. Simulation results for MSE of eigenvalues when there is 20% contamination in high dimensional data set

	Mean					Median					Error				
CPCA	151,618	0,249	0,052	0,007	0,006	150,766	0,169	0,03	0,003	0,003	1,498	0,023	0,006	0,001	0,0007
ROBPCA	0,0948	0,029	0,008	0,013	0,004	0,0208	0,012	0,004	0,009	0,003	0,015	0,004	0,001	0,001	0,0006
OGK	29,8717	1,388	0,769	0,394	0,213	23,4209	1,194	0,716	0,35	0,171	2,086	0,072	0,04	0,022	0,0147
ROGK	0,0801	0,028	0,01	0,018	0,006	0,0267	0,011	0,005	0,014	0,004	0,013	0,005	0,001	0,002	0,0007
rot,OGK	1,3674	0,306	0,103	0,031	0,029	0,4098	0,245	0,071	0,021	0,021	0,231	0,025	0,009	0,003	0,0027
rot,rOGK	0,0515	0,031	0,016	0,025	0,009	0,0262	0,016	0,011	0,021	0,006	0,007	0,006	0,002	0,002	0,0009

Table 10. Simulation results for MSE of eigenvalues when there is 10% contamination in high dimensional data set

	Mean					Median					Error				
CPCA	43,677	0,243	0,045	0,008	0,006	43,426	0,183	0,029	0,003	0,003	0,768	0,02	0,006	0,001	0,0008
ROBPCA	0,0652	0,02	0,012	0,023	0,009	0,0235	0,012	0,006	0,018	0,006	0,009	0,002	0,001	0,002	0,0009
OGK	4,0256	0,906	0,526	0,318	0,256	3,4681	0,848	0,491	0,305	0,23	0,242	0,046	0,026	0,015	0,0166
ROGK	0,0807	0,02	0,01	0,017	0,006	0,0284	0,013	0,006	0,011	0,005	0,012	0,002	0,001	0,002	0,0006
rot,OGK	1,4375	0,493	0,115	0,023	0,008	0,9977	0,348	0,068	0,01	0,003	0,147	0,045	0,015	0,003	0,0014
rot,rOGK	0,0615	0,024	0,017	0,028	0,008	0,0282	0,014	0,012	0,023	0,006	0,008	0,003	0,002	0,002	0,0007

Appendix-2

```
    MATLAB CODE                                   NOTES
function [var,mu]=deviation(x)
med=median(x);
md=mad(x);     # Here, it is also possible to use ‘madc’ function instead of ‘mad’
s=size(x);
Median=(ones(s(1),1))*med;
Mad=(ones(s(1),1))*md;
W=(x-Median)./(Mad);
W=(1 - (W./4.5).^2).^2.*(abs(W)<=4.5);
mu=sum(x.*W)./sum(W);
Mu=(ones(s(1),1))*mu;
rho=((x-Mu)./Mad).^2;                     # First and second steps of the  algorithm
var=((md.^2).*(sum(min(rho,9))))/s(1);
function result =ogk(x)
s=size(x);
[var1,mu1]=deviation(x);
D=diag(sqrt(var1));
y=(inv(D)*x')';                           #Third step of the algorithm
vv=combntns(1:s(2),2);
ss=size(vv);
for i=1:ss(1)
bb{i}=y(:,vv(i,:));
end
for i=1:ss(1)
t(:,i)=bb{i}(:,1)+bb{i}(:,2);
tt(:,i)=bb{i}(:,1)-bb{i}(:,2);
end
[var2,mu2]=deviation(t);
[var3,mu3]=deviation(tt);
U=(var2-var3)/4;                     #Fourth step of the algorithm
UU=zeros(s(2));
for i=1:ss(1)
UU(vv(i,1,:),vv(i,2,:))=U(i);
end
UU=eye(s(2))+UU+UU';
[E,T]=eig(UU);
A=D*E;
z=E'*y';
Z=z';                                      # Fifth step of the algorithm
[var,mu]=deviation(Z);
RO=diag(var);
 v=A*RO*A';
 m=A*mu';
d=sum((((Z-(ones(s(1),1)*mu))./sqrt(ones(s(1),1)*var)).^2)');    #  Sixth  step,     OGK
estimators
do=(chi2inv(0.9,s(2))*median(d))/chi2inv(0.50,s(2));
 w=((d<=do)*1);
 rm=(x'*w')/sum(w);
 dif=x-(ones(s(1),1)*rm');
 rv=(dif'*(diag(w))*dif)/sum(w);
 result=struct('m',{m},'v',{v},'rm',{rm},'rv',{rv});       #Seventh step, R-OGK estimators
```