Selected issues related to online calculation of multivariate robust measures of location and scatter

Daniel Kosiorowski[1], Zygmunt Zawadzki[2]

Abstract

Multivariate location and scatter measures are working horses for a variety of statistical procedures used within the modern Economics. With appearance of new phenomena related to very big data sets, online inference and data processing – a computational complexity of the procedure is pointing on the foreground of scientific researches. Due to existence of outliers in the economic data sets, robust statistical procedures are used more and more often. Unfortunately, a great part of robust estimators of the multivariate location and scatter are computationally and/or memory very intensive and do not allow for the recursive calculation in a similar manner as least squares estimators. In this paper we study possibilities of overcoming these substantial computational difficulties. We focus our attention on three representative estimators: minimum covariance matrix determinant (MCD), Orthogonlized Gnanadeskian/ Kettering estimator (OGK) and the general depth weighted location and scatter estimator (DIS).

Keywords*: multivariate location and scatter; robust procedure; fast algorithm; streaming data*
JEL Classification*:* C53, C55, C14

1. Introduction

The sample mean vector (MV) and the sample covariance matrix (Cls) have been the standard estimators of location and scatter in the multivariate statistics. They are affine invariant and highly efficient for normal models. Moreover, both the MV and the Cls allow for a distributed and recursive calculation what is especially important for the very big data sets or in case of a streaming data analysis (see Anagnostopuoulos et al., 2012).

Unfortunately, economic data sets very often contain outliers or inliers of a various kind, what makes the MV and the Cls useless due to their extreme sensitivity to atypical observations. Although, there are known good robust alternatives for the MV and the Cls, these alternatives are treated as very computationally and/or memory demanding. This computational and/or memory complexity limits an application of the robust measures in data stream analysis or in a mining in very big data sets. The existing algorithms for robust measures calculation are considered as being too complex for online credit card fraud detection, or the financial marketss monitoring.

[1] Corresponding author: Cracow University of Economics, Rakowicka 27, 31-510 Kraków, Polande-mail: daniel.kosiorowski@uek.krakow.pl
[2] Cracow University of Economics, e-mail: zawadzkizygmunt@gmail.com

It is well known that the usual sample mean and sample variance allows for the recursive calculation. Similarly for the sample covariance $c_t = 1/t \sum_{i=1}^{t} (x_i - \bar{x}_t)(y_i - \bar{y}_t)$ calculated from the sequence $(x_1, y_1), (x_2, y_2), \ldots$, we have:

$$t \cdot c_t = (t-1)c_{t-1} + \frac{t-1}{t}(x_t - \bar{x}_{t-1})(y_t - \bar{y}_{t-1}). \quad (1)$$

Equation (1) lead to a naïve recursive algorithm for the Cls. The algorithm can be improved using recursive least squares algorithms which facilitate the revision estimates when new observations became available. Theory of recursive least squares estimation was first explored by Gauss in his original tract on the method of least squares (see Durbin and Koopman, 2001).

A possibility of the recursive calculation of the robust descriptive measures is a subject of very intensive studies nowadays. For a great part of robust measures an answer is negative due to "influential majority" ideas which lie on the ground of the concept of robustness. This "influential majority" may dramatically change with an arrival of new observation.

We understand robustness of the estimator in terms of the influence function (IF) and the finite sample breakdown point (BP) – for further details see Maronna et al. (2006). The BP point serves as a measure of global robustness, while the IF function captures the local robustness.

The rest of the paper is organized as follows: in Section 2, three representative robust location and scatter estimators are briefly described. In Section 3, results of comparisons of the estimators are presented. The paper ends with conclusions, description of the future expected results and references.

2. Robust estimators of location and scatter

The first affine equivariant estimator of multivariate location and scatter which attains a very high BP was the Stahel – Donoho estimator. Nowadays, there are many robust alternatives for the MV and Cls. Many classical high BP point estimators are commonly treated as inefficient at normal populations, and computationally and/ or memory very expensive.

For effective application of the robust location and scatter estimators in streaming data analysis, their recursive and/or distributed formulation are needed (see Muthukrishan, 2006) Below we briefly describe two representative robust estimators and compare them with new robust estimator using so called data depth concept. Further details can be found in Rousseeuw and Van Driessen (1999), Maronna and Zamar (2002) and Todorov and Filzmoser (2009).

2.1. Minimum Covariance Determinant Estimator

For a data set $\mathbf{X}^n = \{X_1, ..., X_n\}$ in $\Re^d$ the *Minimum Covariance Determinant* (MCD) is defined by the subset $\{X_{i_1}, ..., X_{i_h}\}$ of h observations, whose covariance matrix has the smallest determinant among all possible subsets of $\mathbf{X}^n$ size h. The MCD location and scatter estimate T_{MCD} and C_{MCD} are given as the arithmetic mean and a multiple of the sample covariance matrix of the subset:

$$T_{MCD} = \frac{1}{h}\sum_{j=1}^{h} \mathbf{x}_{ij} \tag{2}$$

$$Cov_{MCD} = c_1 c_2 \frac{1}{h-1}\sum_{j=1}^{h} \left(\mathbf{x}_{ij} - T_{MCD}\right)\left(\mathbf{x}_{ij} - T_{MCD}\right)^T \tag{3}$$

where the multiplication factors c_1 (consistency correction factor) and c_2 (small sample correction factor) are selected so that Cov_{MCD} is consistent at the multivariate normal model and unbiased at small samples. A recommendable choice of h is $\lfloor (n+d+1)/2 \rfloor$ because then the BP of the MCD is maximized, here $\lfloor \mathbf{z} \rfloor$ denotes the integer part of $\mathbf{z}$ which is not less than $\mathbf{z}$.

The MCD estimator and all other known affine equivariant high-breakdown point estimators are solutions to a highly non-convex optimization problems. Their computation needs nontrivial algorithms. Due to Rousseeuw and Van Driessen (1999), the fast algorithm for its computation is known. The algorithm is very fast for small data sets but is not feasible for the big data sets. A crucial element comprises of the determinant calculation for subsamples. Using well known LU or Cholesky decomposition relates to complexity of $O(d^3)$ for the determinant and $O(nd^2)$ for the covariance matrix.

2.2. Orthogonlized Gnanadeskian/Kettering estimator

Very often in practice, it suffices for the estimator to be invariant with respect to orthogonal transformations of the data. The scarifying the affine invariance may lead to the improvements in terms of its computational complexity. For a pair of random variables Y_j and Y_k, and a standard one-dimensional dispersion measure $\sigma()$, the covariance c_{ij} between them can be expressed as:

$$c_{jk} = \frac{1}{4}\left(\sigma\left(\frac{Y_j}{\sigma(Y_j)} + \frac{Y_k}{\sigma(Y_k)} \right)^2 - \sigma\left(\frac{Y_j}{\sigma(Y_j)} - \frac{Y_k}{\sigma(Y_k)} \right)^2 \right). \quad (4)$$

Calculation the covariance matrix $[c_{jk}]$ using (4) and robust measure $\sigma()$, e.g., the median of absolute deviation from the median $MAD = Med\left(\left|Z - Med\left(Z\right)\right|\right)$ leads to *Gnanadesikan/Kettenring* scatter estimator. This estimator may not necessary be positive symmetric. Due to Maronna and Zamar (2002) its improved version is known as *Orthogonalized Gnanadesikan/Kettenring* estimator. Its computational complexity in the form (8) for σ taken as the MAD is $O(d^2 n\log(n))$. It is easy to notice that this estimator allows for the distributed and parallel calculation.

2.3. General depth weighted location and scatter estimators

Data depth was originally introduced as a way to generalize the concepts of median and quantiles to the multivariate framework. A depth function $D(\cdot, F)$ associates with any $\mathbf{x} \in \Re^d$ a measure $D(\mathbf{x}, F) \in [0,1]$ of its centrality w.r.t. a probability measure $F \in \mathrm{P}$ over $\Re^d$ or w.r.t. an empirical measure $F_n \in \mathrm{P}$ calculated from a sample $\mathbf{X}^n$. The larger the depth of $\mathbf{x}$, the more central $\mathbf{x}$ is w.r.t. to F or F_n**.** The most celebrated examples of the depth known in the literature are *Tukey* and *Liu* depth (for further details see Zuo, 2004). For our purposes, the most interesting depth seems to be the weighted L^p depth. **The weighted** L^p **depth** $WL^pD(\mathbf{x}; F)$ of a point $\mathbf{x} \in \Re^d, d \geq 1$ being a realization of some d dimensional random vector $\mathbf{X}$ with distribution F, is defined as:

$$WL^pD(\mathbf{x}; F) = \frac{1}{1 + Ew(\left\|\mathbf{x} - \mathbf{X}\right\|_p)}, \quad (5)$$

where w is a suitable weight function on $[0,\infty)$, and $\left\|\cdot\right\|_p$ stands for the L^p norm (when $p=2$ we have usual Euclidean norm). We assume that w is non-decreasing and continuous on $[0,\infty)$ with $w(\infty -) = \infty$, and for $a, b \in \Re^d$ satisfying $w(\left\|a + b\right\|) \leq w(\left\|a\right\|) + w(\left\|b\right\|)$. Examples of the weight functions are: $w(x) = a + bx$, $a, b > 0$ or $w(x) = x^\alpha$. The empirical version of the weighted L^p

depth function is obtained by replacing distribution F of $\mathbf{X}$ in $Ew(\|\mathbf{x}-\mathbf{X}\|_p)=\int w(\|x-t\|_p)dF(t)$ by its empirical counterpart calculated from the sample $\mathbf{X}^n=\{\mathbf{x}_1,...,\mathbf{x}_n\}$

$$WL^pD(\mathbf{z},\mathbf{X}^n)=\left[1+\frac{1}{n}\sum_{i=1}^{n}w\left(\|\mathbf{z}-\mathbf{x}_i\|_p\right)\right]^{-1}. \tag{6}$$

A point for which depth takes the maximum is called the L^p **median** (multivariate location estimator), the set of points for which depth takes value not smaller than $\alpha\in[0,1]$ is multivariate analogue of the quantile and is called the $\alpha-$ central region, $D_\alpha(F)=\{\mathbf{x}\in\Re^d: WL^pD(\mathbf{x},F)\geq\alpha\}$. Fig. 1 presents the L^2 sample depth function contour plot obtained using DepthProc package (see Kosiorowski et al., 2013).

The weighted L^p depth function in a point, has the low BP and unbounded IF. On the other hand, the weighted L^p depth induced medians (multivariate location estimator) are globally robust with the highest BP for any reasonable estimator. The weighted L^p medians are also locally robust with bounded influence functions for suitable weight functions. Unlike other depth functions and multivariate medians, the weighted L^p depth and medians are easy to calculate in high dimensions. The price for this advantage is the lack of affine invariance of the weighted L^p depth and medians, respectively.

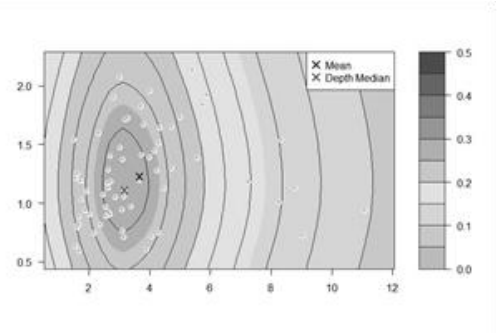

Fig. 1. Sample L^2 depth contour plot (DepthProc package).

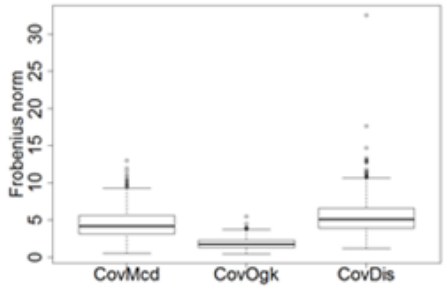

Fig. 2. Boxplots for Fröbenius norm of differences between true and estimated covariance matrices using the MCS, the OGK and the DIS.

Using weighted L^p depth one can define a **depth-weighted mean** with weighted L^p depth:

$$L(F)=\int \mathbf{x}w_1(WL^pD(\mathbf{x},F))dF(\mathbf{x})\Big/ w_1(WL^pD(\mathbf{x},F))dF(\mathbf{x}) \tag{7}$$

Subsequently, a **depth-weighted scatter estimator** based on weighted L^p depth is defined as

$$S(F) = \frac{\int (\mathbf{x} - L(F))(\mathbf{x} - L(F))^T w_2(WL^p D(\mathbf{x},F)) dF(\mathbf{x})}{\int w_2(WL^p D(\mathbf{x},F)) dF(\mathbf{x})}, \tag{8}$$

where $w_2(\cdot)$ is a weight function that can be different from $w_1(\cdot)$.

Note that $L(\cdot)$ and $S(\cdot)$ include multivariate versions of trimmed means and covariance matrices. Their sample counterparts take the forms:

$$T_{WD}(\mathbf{X}^n) = \sum_{i=1}^{n} d_i X_i \Big/ \sum_{i=1}^{n} d_i \tag{9}$$

$$DIS(\mathbf{X}^n) = \sum_{i=1}^{n} d_i \left(\mathbf{X}_i - T_{SD}(\mathbf{X}^n)\right)\left(\mathbf{X}_i - T_{SD}(\mathbf{X}^n)\right)^T \cdot \left[\sum_{i=1}^{n} d_i\right]^{-1}, \tag{10}$$

where d_i are sample depth weights, $w_1(x) = w_2(x) = x$.

Computational complexity of the scatter estimator crucially depend on the complexity of the depth used. For the weighted L^p depth we have $O(d^2 n + n^2 d)$ complexity and good perspective for its distributed calculation (see Zuo, 2004).

3. Comparison of the estimators

In order to compare properties of the estimators we performed simulations using various models differing w.r.t. their dimensionality, and involving departure from i.i.d. setting. Let $\mathbf{X}$ has a distribution F with a location vector $\mathbf{m}$ and a scatter matrix Σ. Let T_{LOC} and T_{SC} denote the location and the scatter estimators calculated from a sample $\mathbf{X}^n \subset \Re^d$. In the simulations, we compared the estimators by means of the E_1 and E_2 criteria defined as:

$$E_1 = \left\| T_{LOC} - \mathbf{m} \right\|_{EUC}, \quad E_2 = \left\| T_{SC} - \Sigma \right\|_{FR} \tag{11}$$

where $\left\|\mathbf{A}\right\|_{FR} = \sum \left|a_{ij}\right|^2 = \sqrt{tr(\mathbf{A}^T\mathbf{A})}$ is the Fröbenius norm , and $\left\|\cdot\right\|_{EUC}$ is the Euclidean norm.

Fig. 2 presents boxplots showing the Fröbenius distances between true population covariance matrix and the estimates obtained by means of the MCD, the OGK and the DIS. Samples consisted of 150 obs. from i.i.d. two-dimensional Student distribution with 3 degree of freedom. In terms of unbiasedness and dispersion, the OGK performed the best, the DIS and the MCD performed similarly. Fig. 3 presents results of 500 simulations from two-regime i.i.d stream

model consisted of two regimes differing w.r.t. their scatter (for the data stream issues see Muthukrishan, 2006; Kosiorowski, 2013).

REG1(90%) +REG2(10%)	REG1(80%) +REG2(20%)	REG1(70%) +REG2(30%)	REG1(60%) +REG2(40%)
MCD $\begin{pmatrix} 6.06 & 3.82 \\ 3.82 & 9.69 \end{pmatrix}$ F=1.2	$\begin{pmatrix} 6.31 & 3.12 \\ 3.12 & 8.63 \end{pmatrix}$ F=2.7	$\begin{pmatrix} 6.61 & 2.37 \\ 2.37 & 7.62 \end{pmatrix}$ F=4.2	$\begin{pmatrix} 6.93 & 1.59 \\ 1.59 & 6.61 \end{pmatrix}$ F=5.7
OGK $\begin{pmatrix} 3.33 & 2.11 \\ 2.11 & 5.35 \end{pmatrix}$ F=6.1	$\begin{pmatrix} 3.43 & 1.73 \\ 1.73 & 4.78 \end{pmatrix}$ F=7.4	$\begin{pmatrix} 3.55 & 1.32 \\ 1.32 & 4.21 \end{pmatrix}$ F=8.1	$\begin{pmatrix} 3.69 & 0.91 \\ 0.91 & 3.64 \end{pmatrix}$ F=8.8
DIC $\begin{pmatrix} 6.70 & 3.45 \\ 3.45 & 9.83 \end{pmatrix}$ F=1.7	$\begin{pmatrix} 7.05 & 2.88 \\ 2.88 & 9.13 \end{pmatrix}$ F=2.9	$\begin{pmatrix} 7.41 & 1.73 \\ 1.73 & 8.39 \end{pmatrix}$ F=4.0	$\begin{pmatrix} 7.69 & 1.73 \\ 1.73 & 7.62 \end{pmatrix}$ F=5.1
Cls $\begin{pmatrix} 12.2 & 7.47 \\ 7.4 & 18.45 \end{pmatrix}$ F=10.9	$\begin{pmatrix} 12.50 & 6.25 \\ 6.25 & 15.76 \end{pmatrix}$ F=9.7	$\begin{pmatrix} 13.5 & 5.21 \\ 5.21 & 15.76 \end{pmatrix}$ F= 9.7	$\begin{pmatrix} 13.85 & 3.94 \\ 3.94 & 14.21 \end{pmatrix}$ F=8.7

Table 1 Performance of the MCD, the OGK, the DIS and the CLS scatter estimators.

We analyzed a performance of the MCD, the OGK, the DIS, and the Cls by calculating them from moving windows consisting 90% ,80%, 70%, 60% obs. from the first regime, and 10%, 20%, 30%, 40% obs. from the second regime of the stream. The stream was generated from two i.i.d. 2D Student distributions with 3 degree of freedom with the same location but differing w.r.t. the scatter matrices:

$$S_1 = \begin{bmatrix} 6 & 4.5 \\ 4.5 & 10.5 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 6 & -1 \\ -1 & 2 \end{bmatrix}.$$

The results are presented in tab. 1. Above each of the matrices, the mean Fröbenius distance between estimate and the dispersion matrix of the first regime is placed (F).

For MCD, OGK, and DIS we observed similar directions of the changes in the estimates values. The changes of the DIS estimator were "more discontinuous" manifesting in increase of the mean Fröbenius distance from 2.86 to 4. The smaller BP of the DIS estimator, seems to be here an advantage. The performance of Cls in the considered situation was non-informative. In case of the location estimator, both L^p median and proposed estimator (10) outperformed the MCD in terms of a computation time. A memory complexity of the OGK without the orthogonalization does not exceed $O(n+d^2)$. For the MCD we obtain $O(k(h+d^2))$, where h

denotes subsample length, k denotes the number of starting subsamples. Finally for the DIS, its memory complexity equals $O(n+d^2)$. The memory complexity of L^p depth equals only $O(1)$.

In order to investigate a performance of the estimators we used real financial data consisted of high-frequency quotations of 20 companies belonging to Dow Jones Industrial index. The considered dataset involved a period from April 2008 to June 2013.

In the first part of the study we calculated time of the algorithm execution for data sets differing with respect to their size. Time of calculation of the OGK and the MCD estimators even for a data set consisting of 100000 observations was relatively short. It is worth noticing that MCD estimator is much faster than OGK estimator, which is by default used for very big data sets within "reference" package {rrcov}. Calculation of the DIS estimator is much slower than for the MCD and OGK estimators.

In order to check if the considered estimators lead to similar results we calculated the covariance matrices for data sets representing consecutive months (63 months), and next we calculated Fröbenius distances between them. Fig. 3 presents the results. Apart from a period from the four quarter of 2008 to the second quarter of 2009 – all the estimators indicated the similar scatter. Further we compared the MCD, the OGK and the DIS with the classical algorithm of the Cls estimation. The results are presented on the Fig. 4. The closest results to the Cls were obtained using the OGK. We observed significant deviations in case of the DIS. The distance between the OGK and the MCD to the Cls retained on the constant level but the distance between the DIS and the Cls differed significantly. In the period of the crisis of 2008 year, the DIS estimator significantly departed from the rest. It is a very interesting result because in this period started the financial crisis. Fig. 5 presents the correlation matrices visualization prepared basing on the MCD, OGK and DIS estimates. The DIS based correlation matrix indicates weaker correlations for the particular instruments than the OGK and the MCD estimators.

Conclusions

A great part of multivariate robust measures is not computationally feasible for the real application in the online Economy. The analysis of the empirical example results to a conclusion that the DIS cannot compete with the OGK and the MCD in the context of analysis of very big data sets. The DIS enables however for the parallel calculation, what can improve its properties. Both the MCD and the OGK however enable for the parallel calculation too. We can recommend

however the MCD and the OGK for data stream analysis. A calculation of the covariance matrix for 20 stocks quoted every 5 minutes in a 5 years period takes about 5 seconds on the standard PC produced in 2013 year.

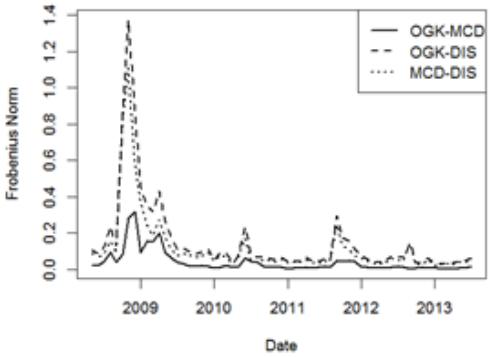

Fig. 3. The Fröbenius distance between the scatter matrices calculated for 63 consecutive months

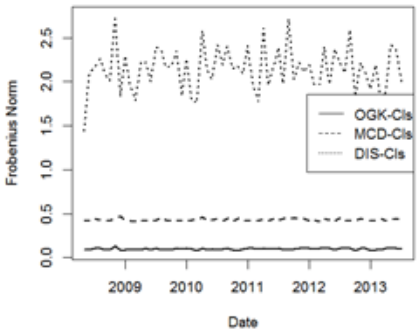

Fig. 4. The Fröbenius distance between the OGK, MCD the DIS scatter matrices calculated for 63 consecutive months and the Cls

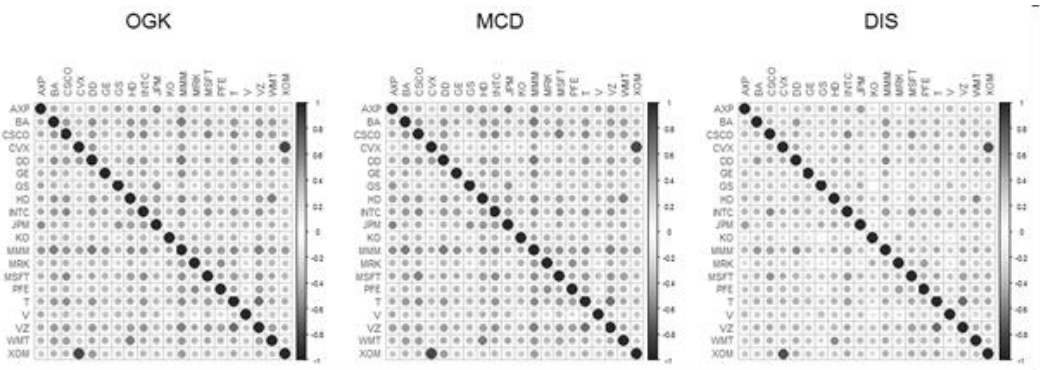

Fig. 5. Visualisation of the OGK, the MCD and the DIS estimates for 20 stocks from Dow Jones Industrial in the September of 2008 year.

The observed behavior of the DIS in the beginning of the financial crisis of the 2008 year seems to be especially interesting. The DIS is robust but is not very robust in terms of the BP. The DIS alerts us earlier as to the scatter of the stream change than the MCD and the OGK.

The presented in this paper estimators do not allow for their recursive calculations due to a fact that they critically depend on the influential majority of the data. This majority can dramatically change with an arrival of new observation.

Acknowledgments

The authors thank for financial support from Polish National Science Center grant UMO-2011/03/B/HS4/01138.

References

Anagnostopuoulos, Ch., Tasoulis, D. K., Adams, N. M., Pavlidis, N. G., & Hand, D. J. (2012), Online linear and quadratic discriminant analysis with adaptive forgetting for streaming classification, *SADM*, 5, 139-166.

Devlin, S. J., Gnanadesikan, R., & Kettenring, R. J. (1982), Robust estimation of dispersion matrices and principal components, *JASA*, 76, 354-362.

Durbin, J., & Koopman, S. J. (2001), *Time series analysis by state space methods*, Oxford Statistical Science Series 24, Oxford

Kawahara, Y., & Sugiyama, M. (2012), Sequential change-point detection based on direct density ratio estimation, *SADM*, 5, 2012, 114-127, doi:10.1002/sam.10124.

Kosiorowski, D., Zawadzki, Z., Bocian, M., & Wegrzynkiewicz, A. (2013), Depthproc package in multivariate time series mining, *Acta Universitatis Lodziensis – FO*, 286, 243-251.

Maronna, R. A., Martin, R. D. & Yohai, V. J. (2006), *Robust statistics – theory and methods*, Chichester: Wiley.

Maronna, R. A., & Zamar, R. H., (2002), Robust estimation of location and dispersion for high-dimensional datasets, *Technometrics*, 44, 307-317.

Muthukrishan, S. (2006), *Data streams: algorithms and applications*, Now Publishers.

Rousseeuw, P. J., & Van Driessen, K. (1999), A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, 41, 212-223.

Todorov, V., & Filzmoser, P., (2009), An object oriented framevork for robust multivariate analysis, *JSS*, 32(3), 1-47.

Zuo, Y., (2004), Robustness of weighted L^p – depth and L^p median," *ASA*, 88, 215-234.

Zuo, Y., Cui, H., & He, X. (2004), On the Stahel-Donoho estimator and depth weighted means for multivariate data. *Ann. Statist*. 32, 167-188.