Multivariate median

Yangming Li

April 2019

Abstract

Multivariate median is an important concept in the robust statistics. This project reviews the concept of three multivariate medians: vector of margin of median, oja median and the spatial median introduced in the project 1. Concepts of two new medians: Tukey median and Liu's median are introduced in this new project. Properties of them like robustness, efficiency are discussed. Simulation studies, visualization about the data depth are carried out.

Keywords：

Clustering, Tukey median, Liu's median, Data Depth, Efficiency, Bag Plot, Contour Plot.

1. Review multivariate median in project 1

In project 1, we introduced three different multivariate medians: vector of margin of median, spatial median, and oja median. Let $X_1, \ldots, X_n$ denotes a sample of p dimensional random variables from a distribution F, the multivariate median $T(X)$ is a minimization of the criterion function $D(t)$ with respect to t. Marginal median is a minimizer for the sum of Manhattan distance, spatial median is a point which sum of samples' Euclidean norm is smallest. In $\mathbb{R}^p$, for oja median, the sum of the volume of all p-variate simplix is minimized. We introduce the method to compute spatial median by Weiszfeld's algorithm and compute oja median by Grid-based Algorithm. We introduce the concept of marginal sign, spatial sign, and oja sign and we test the robustness of these three medians based on the sign and sign covariance matrix .Moreover, we detected the outlier by the Sign Covariance Matrices. Marginal median and spatial median are robust estimator with breakdown point 1/2, However, they are not affine equivariant. Oja median has 0 breakdown

point, but it is affine equivariant.

The rest of the project 2 is organized as follows. In Section 2, we find a new algorithm based on spatial sign and median to cluster the data. In Section3, we talked about the Tukey Median and the Liu's simplicial depth Median, the robustness of these medians and the finite sample efficiency of these medians. In Section 4, we introduce the concept of the depth and visualize the depth by the contour plot and the bag plot.

2. Clustering based on Spatial sign and Spatial median

One of useful application of multivariate median in the high dimension case is to enhance the robustness of kmeans clustering method by J MacQueen(1967). Kmeans clustering is not robust since kmeans use the mean to calculate the centroid of the clusters, the mean is not robust and have a low breakdown point which is $\frac{1}{n}$, where n is the number of points in the data set, so kmeans algorithm will be easily influenced by the outliers. To improve this, We try to cluster the data by the spatial sign and calculate the centroid by the Spatial median. As we talked in project 1 the Spatial sign is $\mathrm{sgn}(t) = S_2(t) = \begin{cases} \frac{t}{\|t\|}, & \text{if } t \neq 0 \\ 0, & \text{if } t = 0 \end{cases}$ we all known the spatial sign is robust against extreme value. So in this case , we replace original data set by the Spatial sign of the data set , and we perform data by clustering the spatial sign, Moreover, the spatial median has high breakdown value 0.5, which is way much robust than the mean. We hope to use spatial median's robustness to improve the robustness of the clustering algorithm. Kmeans algorithm iterations parts have two steps calculations, one is the assignment steps: assign the points into the nearest centroid step, the other one is recompute centroid step, in kmeans algorithm, it just uses the mean function to compute the average of the points in each cluster, for point x_i assigned to center μ_j, we compute μ_j as $\mu_j = \frac{1}{n_j}\sum_{x_i \to \mu_j} x_i$, for our new algorithm, we use the spatial median

which is $\mu_t = \text{argmin} \sum_{x_i \to \mu_t} ||x_i - \mu_t||$, the extra part of the kmeans algorithm remains the same. We use the 5-dimensional dataset "woodmod.dat" from "robust" package to test our algorithm. it is known to have 4 outliers in it. Figure 1 (b)shows that since we have 4 outliers, these outliers are clustered incorrectly to the wrong groups. 1(c) shows Our new algorithm clustered them into the correct groups.

(a) (b) (c)

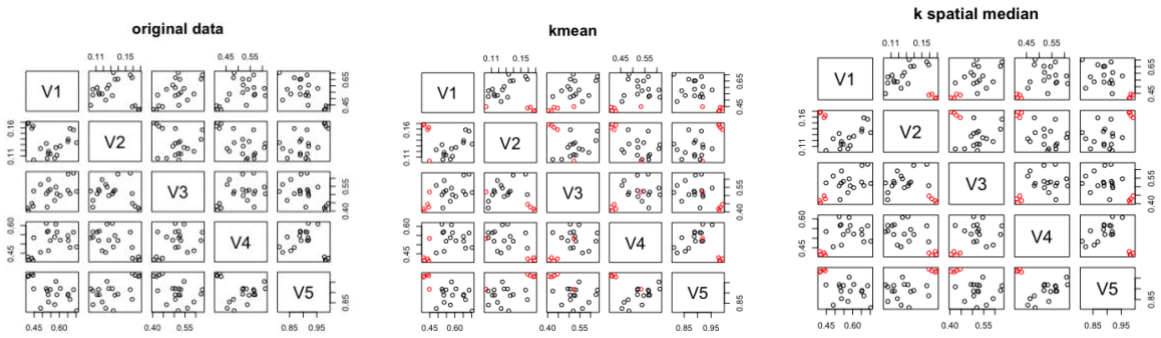

Figure1: scatter matrix plots of the clustering method using the dataset "woodmod", express two different clusters by two different colors (red, black): (a) original data set, (b) the data set clustered by the kmeans algorithm. (c) the data set clustered by the spatial median.

3. Introduction to the two new medians

3.1 The Tukey Median and the Liu's Simplicial Depth Median

We are going to introduce to another two new medians: Turkey and simplicial depth median. Tukey (1975) firstly mentioned the concept of the Tukey median. Similar to the oja median, the turkey median is the point which have their maximum corresponding depth. The turkey median is related to the concept of the half space depth. By Hodges (1955), Tukey (1975), the half space depth is $\text{hdepth}(\theta, x_n) = min\#\{i; u^T x_i \geq u^T \theta\}$. It means for a query point with respect to a set S of n points in R^d. For each closed half space that contains count the number of points in S which are in the half space. Take the minimum number found over all half spaces to be the depth

of θ . Aloupis (2001) showed the simple example in R^2,see figure 2(a), place a line through so that the number of points on one side of the line is minimized. For point A the minimum number found on the left of the line is 1, for B are 3 points, for C are 4 points. As a result, the half space depth for A is 1, for B is 3, and for C is 4 in this case. Computation of the tukey median is just find the point has the maximum value of the minimized point found on the left of the line.

For the simplicial median, Liu (1990) defined this idea. The simplicial median in R^d is a point in R^d ,which contained in the most simplices formed by subsets of d+1 data points. Aloupis, (2001) used the 4 points in R^2 case to illustrate this idea. As can be seen from the figure2(b). We should find the point which is contained by most of triangles. The straight computation straight forward method of the simplical median in R^2 is to partition the plane into cells which have segments between points as boundaries. In the Aloupis (2001) example showed in figure2(c), we have A,B,C,D,E,F,H,G,I cells. First, notice that every point within a cell has equal simplicial depth concept which will be introduced in depth section, Furthermore, a point on a boundary between two cells must have depth at least as much as any adjacent interior points. Similarly, an intersection point (where more than two cells intersect) must have depth at least as much as any adjacent boundary point. So we know that the maximum value, must in the intersection point. Therefore, by determining how many triangles contain each intersection point, we can evaluate it and find the simplicial median.

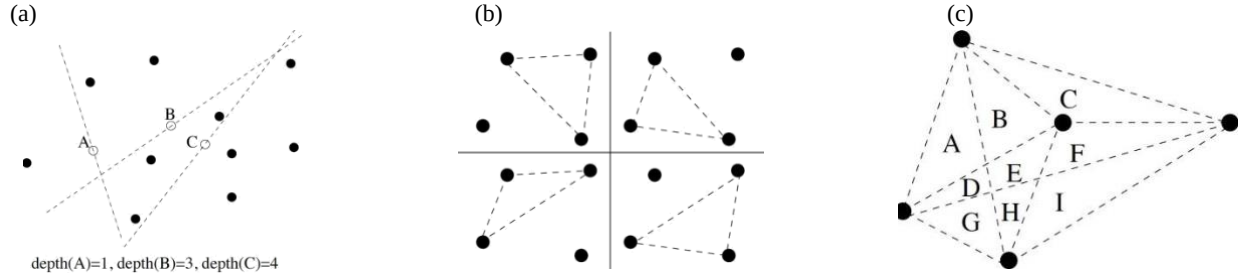

Figure2: simplicial examples for the turkey median, simplicial median (a) the tukey median example ,(b) the simplical depth example, (c) the computation of simplical depth example.

3.2.Robustness and breakdown properties of three multivariate medians(Oja,Chen,Liu)

Chen (1995) proved Tukey's half-space depth-based median is known to possess an asymptotic breakdown of $\frac{1}{3}$ when the underlying probability distribution is absolutely continuous and angularly symmetric.

Oja et al. (1990) showed that it is possible to break oja's simplicial volume based median by corrupting only two data points, and consequently, it has a very poor 0 asymptotic breakdown point. However, for the finite sample breakdown point, Niinimma (1995) showed Oja median is special which breakdown point depends on the dispersion of the corrupted points. If in this bivariate case all the corrupted points are identical (the dispersion is minimal) then the breakdown point is $\frac{1}{3}$. it highly depended on dispersion of contaminated data point will changed by different dispersion of data, this minimal dispersion case is $\frac{1}{3}$ .If the dispersion of contaminated data is not allowed to exceed the dispersion of the original data set, then the breakdown is at least 0. 293.In the general k-variate case the breakdown point is $\frac{1}{k}$.

Chen (1995) observed that when the underlying distribution is absolutely continuous, an upper bound for the asymptotic breakdown point of Liu's simplicial depth-based median is $\frac{1}{d+2}$, where d is the dimension of the data.

We use three examples to compare the robustness of these three medians: we generating 50 observations from standard bivariate normal distribution with $\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$. compute the liu's simplicial depth median and tukey's half-space depth median and oja median. As the plot (a) in figure 3 shows: there are very little visible differences among the positions of these three

multivariate medians. We generate the same 50 bivariate normal observations with another 25 observations from bivariate normal with $\mathcal{N}\left(\begin{bmatrix}10\\10\end{bmatrix},\begin{bmatrix}0.01 & 0\\0 & 0.01\end{bmatrix}\right)$. As can be seen from the figure2(b), Tukey's half-space depth median is outside the original data cloud. So turkey median is easily influence by changing the distribution of some points. In this case we just change the $\frac{1}{3}$ of the total data points. It is consistent with the asymptotic breakdown $\frac{1}{3}$ property we stated before.Oja median is not robust and shifted. Then we added 7 replications of the point (10,10) to the set of 50 original observations. As can be seen from the figure 3(c), we have observed that one can break Liu's median by just replicating some corrupted observation only a few times

(a) (b) (c).

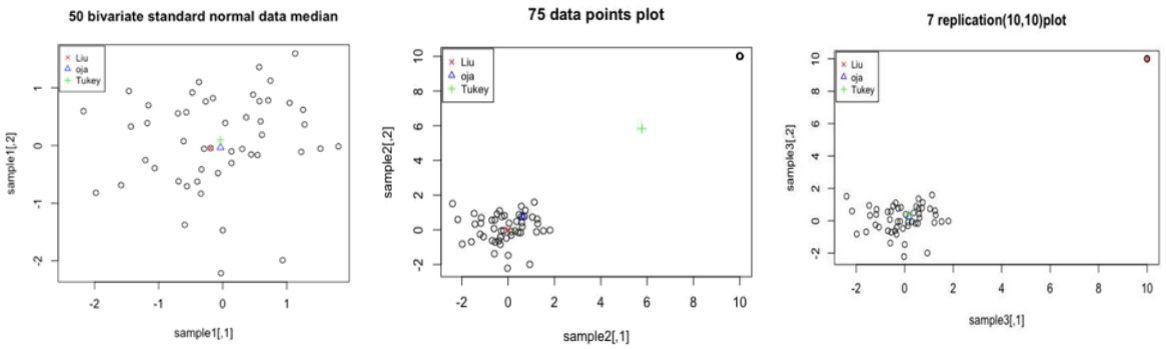

Figure 3: different bivariate medians for corrupted and uncorrupted data: (a) 50 observations from standard bivariate normal distribution with $\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$. (b) 50 bivariate normal observations with another 25 observations from bivariate normal with $\mathcal{N}\left(\begin{bmatrix}10\\10\end{bmatrix},\begin{bmatrix}0.01 & 0\\0 & 0.01\end{bmatrix}\right)$.(c) added 7 replications of the point (10,10) to the set of 50 original observation

3.3Finite-Sample Efficiency of three medians

We want to use the simulations to compare the finite sample efficiency of these three medians. For several sample size n =80, 90,100, generated m=500 samples called T_j^* from the standard normal distribution to calculated the oja median, the tukey median and the Liu's median. Take the Tukey median for example, for the m Tukey medians $T_1^*, \ldots, T_m^*$ we compute the bias, the empirical covariance matrix and the empirical efficiency. The results are shown in Table.1. The

first column contains sample size n. The next two columns show the coordinates of the bias $(\text{bias1}, \text{bias2}) = \overline{T^*}=\underbrace{\text{average}}_{j=1\ldots mn} T_j^*$. Empirical variance-covariance matrix $\hat{A}=\frac{1}{m-1}\sum_{j=1}^{m}(T_j^* - \overline{T^*})(T_j^* - \overline{T^*})^t$. The final column is the empirical (Monte Carlo) efficiency, computed as $\text{eff} = \frac{1}{n\sqrt{\det(\hat{A})}}$, Table 1 shows that the empirical efficiency of the tukey median is in the range of (62%-76%), the Table 2 shows the empirical efficiency of the oja median is around (54%-67%),the Table 3 shows the empirical efficiency of the liu's median is around (35%-43%). So, we can conclude that for the empirical efficiency the tukey median>oja median>liu's median,tukey median has the smallest bias in this case.

Table1:bias and effiency of tukey median **Table2**:bias and efficiency of oja median **Table3**:bias and efficicency of Liu's median

n	bias	eff
80	-0.005, -0.004	0.625
90	-0.001 ,0.004	0.744
100	0.000,-0.001	0.765

n	bias	eff
80	-0.006, -0.004	0.546
90	0.001 ,0.004	0.614
100	-0.001,-0.004	0.676

n	bias	eff
80	0.008, 0.006	0.358
90	0.002 ,0.006	0.407
100	0.004,-0.005	0.434

4. Data depth concept and visualization

Since multivariate median is highly related to the data depth, data depth is an important concept for the multivariate median. In general, a data depth is a way of measuring how deep a given point x∈ R^d with respect to a given data cloud $\{x_1, x_2, \ldots, , x_n\}$. It is a good way to order the data in R^d, multivariate median always defined as the maximum value of the depth. We introduce three types of depth related to the oja, tukey and liu's median. By oja (1983), the oja median has the maxmum oja depth. Oja depth is defined as the $Odepth = (1 + \sum_{(i_1,\ldots i_d)}\{volume\ S[\theta, x_{i_1}, \ldots, x_{i_d}])^{-1}$. it is just the summation of the simplex

formed by the data points. By Hodges (1955), Tukey (1975) , the half space depth is hdepth$(\theta, x_n) = min\#\{i; u^T x_i \geq u^T\theta\}$ which is talked about in section before. The simplicial depth is defined by liu (1990), which is $sdepth(\theta, x_n) = \#\{(i_1, i_2, .. i_{d+1});\ \theta \in S[x_{i1}, x_{i2}, \ldots, x_{id+1}]\}$,where $S[x_{i1}, x_{i2}, \ldots, x_{id+1}]$ is the closed simplex with vertices $x_{i1}, x_{i2}, \ldots, x_{id+1}$.

4.1 Visualization of depth median by bag plot

The bagplot was proposed by Rousseeuw et al. (1999) and can be considered a generalization of the famous univariate boxplot. This data depth representation is constructed using the location depth measure. The main part of the bagplot is the bag which contains 50% of the observations, and within the bag the middle red point is the Tukey median as showed in figure5, the observation with the maximal depth. This graphic is also composed by a fence, which separates the outliers from the other observations, and a loop, where lie on the observations, that do not belong to the bag but are inside the fence. the bagplot shows several characteristics of the data: location of data (the depth median), spread of the data (the size of the bag),the correlation of the data(the orientation of the bag), the skewness of the data(the shape of the bag and the loop), and the tails of the data (the points near the boundary of the loop and the ouliers). The points outside the fence are flagged as outliers. We plot the bag plot of the same three data set described before: 50 observations from standard bivariate normal distribution with $\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$, the same 50 bivariate normal observations with another 25 observations from bivariate normal with $\mathcal{N}\left(\begin{bmatrix}10\\10\end{bmatrix}, \begin{bmatrix}0.01 & 0\\0 & 0.01\end{bmatrix}\right)$, added 7 replications of the point (10,10) to the set of 50 original observations. In figure 5(a), we can clearly see three outliers stay outside the fence. In figure

5(b), we can clearly see the tukey median is shifted and the size shape of the bag are changed, and in figure 5(c), the 7 replication is identified as an outlier in the box plot.

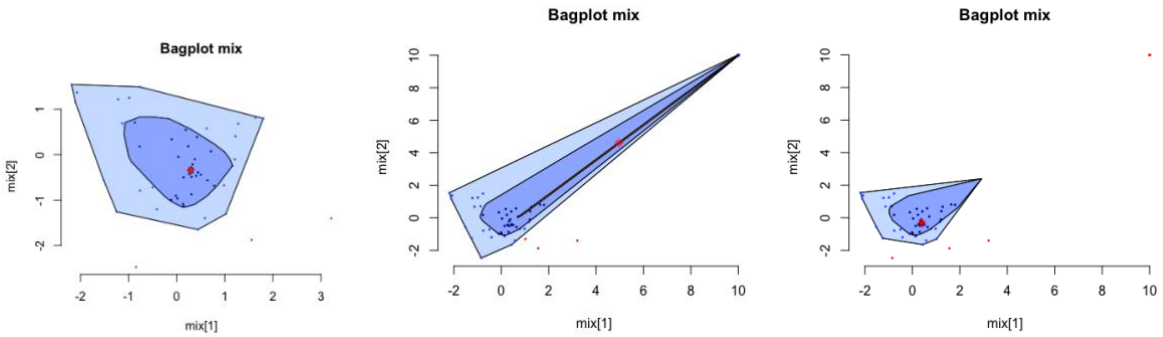

Figure 5: bag plot of the corrupted and uncorrupted data : (a) 50 observations from standard bivariate normal distribution with $\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$. (b) 50 bivariate normal observations with another 25 observations from bivariate normal with $\mathcal{N}\left(\begin{bmatrix}10\\10\end{bmatrix},\begin{bmatrix}0.01 & 0\\0 & 0.01\end{bmatrix}\right)$.(c) added 7 replications of the point (10,10) to the set of 50 original observation

4.2 visualization of depth by contour plot

By (Tukey, 1977), Depth concept is related to rank of the data. When we rank data points in one dimension, we give the extreme points the depth=1. The extreme points can be lowest value or highest value. The data values with the second lowest and second highest rank have depth=2, and so on. Since the median is half-way the ordered list, the median will be the point with maximal depth. For visualize the ranking of the depth, depth contour is a useful tool, which means a line connecting points of equal depth below the hydrographic datum. As Ruts, I., & Rousseeuw (1996) defined the contour concept : $D_k = \{x \in \mathrm{R}^p; \mathrm{depth}(x; X) \geq k\}$. D_k is called the contour of depth k, interior points of D_k have depth at least k, and the boundary points have depth equal to k. The different depth contours form a nested sequence, because D_{k+1} is contained in D_k. The outside contour is the convex hull. We plot the contours based on simplicial depth and half space depth for 50 observations from standard bivariate normal distribution with $\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$. As can be seen from the figure 6 plot(a,b), the shape of the contours are different, for simplicial

depth, the outter contour are not totally convex and smooth, compared to the contour plot by the halfspace depth, then we plot the contour for the 50 bivariate normal original observations with another 25 observations from bivariate normal with $\mathcal{N}\left(\begin{bmatrix}10\\10\end{bmatrix}, \begin{bmatrix}0.01 & 0\\0 & 0.01\end{bmatrix}\right)$ with the half space depth. We find the inner contours are almost not influenced by the outliers, which means the deeper the contour, the more robust it is with respect to outliers in the point cloud

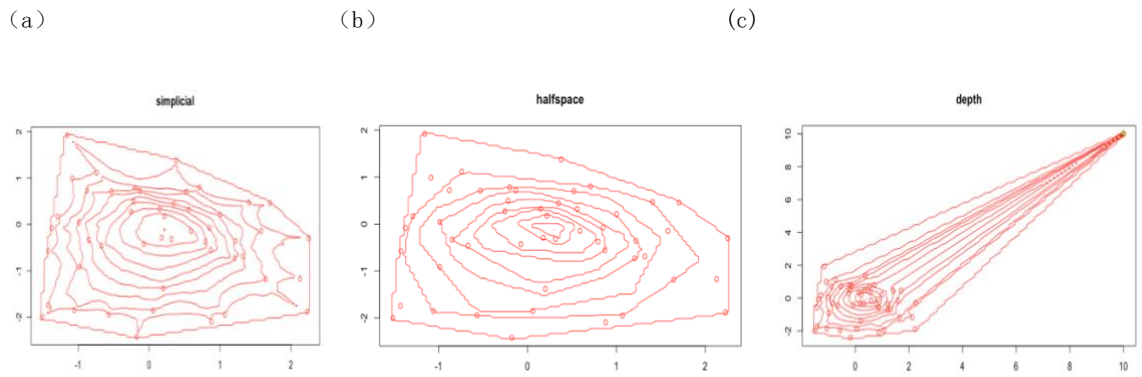

Figure 6: visualize the contour plot (a) contour plot of 50 observations $\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$by simplicial depth, (b) contour plot of 50 observations $\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$,by halfspace depth. (c) 50 bivariate normal original observations with another 25 observations from bivariate normal with $\mathcal{N}\left(\begin{bmatrix}10\\10\end{bmatrix}, \begin{bmatrix}0.01 & 0\\0 & 0.01\end{bmatrix}\right)$

5. Conclusions and future application

In this project, we introduce a new way to clustering the data using the spatial sign and spatial median. It improves the robustness of the clustering. The two new median we introduced has both unique breakdown properties. Turkey median is easily influence by changing the distribution of some number of some points. Liu's median can be breaked by just replicating some corrupted observation only a few times. Finally, bag plot and contour plot are good ways to visualize the data depth, For the future application, Wolfe, Mannos (1979) shows Tukey median filter is widely used in image processing for applications ranging from noise reduction to dropped line replacement. Jörnsten & Zhang (2002) developed a cluster validation and

visualization tool based on the within-cluster data depths, and the cluster data depths with respect to competing clusters.

reference

Aloupis, G. (2001). On computing geometric estimators of location (Doctoral dissertation, McGill University Libraries).

Chen, Z. (1995). Robustness of the half-space median. Journal of statistical planning and inference, 46(2), 175-181.

Chen, Z. Q. (1995). Bounds for the breakdown point of the simplicial median. Journal of Multivariate Analysis, 55(1), 1-13.

Eddy, W. F. (1977). A new convex hull algorithm for planar sets. ACM Trans. Math. Softw., 3(4), 398-403.

Hodegs J. Ž (1955). A bivariate sign test. Ann. Math. Statist. 26 523527.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. The Annals of Statistics, 18(1), 405-414.

Niinimaa, A., & Oja, H. (1995). On the influence functions of certain bivariate medians. Journal of the Royal Statistical Society: Series B (Methodological), 57(3), 565-574.

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297)

Oja, H. (1983). Descriptive statistics for multivariate distributions. Statistics & Probability Letters, 1(6), 327-332.

Jörnsten, R., Vardi, Y., & Zhang, C. H. (2002). A robust clustering method and visualization tool based on data depth. In Statistical Data Analysis Based on the L1-norm and Related Methods (pp. 353-366). Birkhäuser, Basel.

Ruts, I., & Rousseeuw, P. J. (1996). Computing depth contours of bivariate point clouds. Computational Statistics & Data Analysis, 23(1), 153-168.

Tukey, J. W. (1975). Mathematics and the picturing of data. In: James, R. D. (ed.) Proceedings of the International Congress of Mathematicians (Volume 2) 523{531, Canadian Mathematical Congress.

Wolfe, G. J., & Mannos, J. L. (1979, December). Fast median filter implementation. In Applications of Digital Image Processing III (Vol. 207, pp. 154-161). International Society for Optics and Photonics.

Rousseeuw, P. J., Ruts, I., & Tukey, J. W. (1999). The bagplot: a bivariate boxplot. The American Statistician, 53(4), 382-387.

Tukey, J. W. (1975). Mathematics and the picturing of data. In: James, R. D. (ed.) Proceedings of the International Congress of Mathematicians (Volume 2) 523{531, Canadian Mathematical Congress.