Available online at www.alphanumericjournal.com

alphanumeric journal

The Journal of Operations Research, Statistics, Econometrics and Management Information Systems

Volume 10, Issue 1, 2022

Received: December 23, 2021
Accepted: June 15, 2022
Published Online: June 30, 2022

AJ ID: 2022.10.01.STAT.02
DOI: 10.17093/alphanumeric.1041138
Research Article

Teaching the Median with Terms of Absolute Value, Differentiability, and Optimization

Mehmet Hakan Satman, Ph.D.

Prof., Department of Econometrics, Faculty of Economics, Istanbul University, Istanbul Türkiye, mhsatman@istanbul.edu.tr

* İstanbul Üniversitesi İktisat Fakültesi Ekonometri Bölümü Rektörlük Merkez Bina Beyazıt, Fatih İstanbul Türkiye

ABSTRACT

The verbal definition of the sample median sounds a bit strange in the early Statistics courses in the context of being non-mathematical or non-functional. It is also interesting that estimators based on an analytical calculation such as the sample mean are equally strange, but not seen as strange by the students as the median estimator. In this study, we have expanded the studies on teaching the sample median with its optimization definitions. We have also shown that such definitions provide a natural way of understanding the sample median in multivariate case and regression analysis. Seeing that statistical estimators, from the simplest to the most complex, are obtained as a solution to an optimization problem can pave the way for other types of insights.

Keywords: Teaching, Statistics, Absolute Value, Median, Optimization

1. Introduction

Teaching median is easy but it is quite difficult to have a comprehensive understanding without using the terms absolute value, differentiability, and optimization tools. In the univariate case, the sample median is the observation in the middle of the ordered data when the number of observations is odd. It is the average of the two values in the middle of the ordered observations if the number of observations is even. On the other hand, because the arithmetic mean has an algebraic definition, these two descriptive tools are considered to have very different concepts. On the contrary, these two estimators consist of solving similar optimization problems [2, 5].

Besides the sample median and mean, many statistical estimators have optimization based definitions: the sample mod maximizes the frequency, ordinary least squares estimator minimizes the sum of squared residuals, explicit optimizers such as Maximum-Likelihood estimators always maximize an objective function, K-means minimizes within-group variances and maximizes between-group variance, etc. Explaining the basic statistical estimators with their optimization definitions facilitates the transition to extensions and multi-dimensional versions of these estimators in a uniform way.

In Section 2, the optimization based definitions of sample mean and median are introduced. Since the absolute value function is not differentiable in all real points, computational performance of a smooth version is compared with the absolute value function. In Section 3, it is shown that how the optimization-based approach can easily be generalized to multivariate case. In Section 4, a median based robust regression estimator, LMS, is expressed in terms of absolute value and optimization. In Section 5, it is shown that the LAD estimator shares the similar solution technique with the sample median in univariate case. Finally, in Section 6, we conclude.

2. The univariate case

Suppose $x = (x_1, x_2, \dots, x_n)$ is a vector of n observed values. The sample mean $\bar{x}$ minimizes

$$L = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Since L is in quadratic form, the existence of a global minimum is always guaranteed. It can be proved by taking the derivative, equating it to zero, and solving the equation

$$\frac{dL}{d\bar{x}} = \sum_{i=1}^{n} -2(x_i - \bar{x}) = 0$$

yields

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.

Since L is proportional to sample variance, the sample mean also minimizes the sample variance, by definition.

The sample median, b, is generally presented as an order statistic, that is, it is the value at the middle of the ordered observations in univariate case. In other terms, it is expected that the half of the observations are less than the sample median whereas the remaining ones are greater than the sample median. That implies that

$$L_b = \sum_{i=1}^{n} |x_i - b| \tag{1}$$

is minimized where | . | is the absolute value function. L_b is a continuous function in $\mathbb{R}$ whereas its first derivative

$$\frac{d|x|}{dx} = \frac{x}{|x|}$$

is not defined for $x = 0$. The derivative can also be expressed as a truncated signum function

$$\frac{d|x|}{dx} = \begin{cases} -1 & , x < 0 \\ 1 & , x > 0 \end{cases}$$

where $x \neq 0$. The absence of derivative for $x = 0$ does indeed pose a real trouble when the number of observations is odd as at least one observation is equal to sample median in this case.

Smoothing the absolute value function can be used to solve this problem. Replacing the absolute value by

$$\sqrt{x^2}$$

yields exactly the same results with the absolute value function, however, the derivative

$$\frac{d\sqrt{x^2}}{dx} = \frac{1}{2\sqrt{x^2}}$$

is not defined for $x = 0$. Another approximation

$$\lambda(x) = \frac{xe^{kx} - xe^{-kx}}{e^{kx} + e^{-kx}}$$

fits well with the absolute value for large values of k [6]. The first derivate of $\lambda(x)$ is

$$\frac{d\lambda(x)}{dx} = \frac{4kxe^{2kx} + e^{4kx} - 1}{(e^{2kx} + 1)^2}$$

and it is defined for $x \in \mathbb{R}$. Since the function and its first derivative is continuous, any gradient based optimization method can be used to find a solution for L_b, such as gradient descent, possibly with hand calculations. Assume that b_0 is the initial solution for L_b. Then,

$$b_1 = b_0 - \alpha \frac{d\lambda}{dx}$$

is the next step estimate of sample median. After iterating many steps, it is expected b_t to converge to sample median, where t is the number of iterations.

Interestingly, the sample median b defined in Equation (1) can also be expressed as the solution of a goal programming problem, a special member of linear programming. Suppose $u_1^- > 0$ if $x_1 - b < 0$ and $u_1^+ > 0$ if $x_1 - b > 0$. When $u_1^- = u_1^+ = 0, x_1 = b$. By these definitions, negative and positive deviations from the median b are expressed using non-negative terms. Minimizing sum of the deviations yields the objective function

$$\min z = u_1^- + u_1^+ + u_2^- + u_2^+ + \cdots + u_n^- + u_n^+$$

subject to the constraints

$$x_1 - b + u_1^- - u_1^+ = 0$$

$$x_2 - b + u_2^- - u_2^+ = 0$$

$$\vdots$$

$$x_n - b + u_n^- - u_n^+ = 0$$

where

$$u_1^-, u_1^+, u_2^-, u_2^+, \ldots, u_n^-, u_n^+ \geq 0$$

$b \in \mathbb{R}$.

Expressing negative and positive deviations from the sample median replaces the absolute value function. The latest form of the problem is a standard linear programming problem with n constraints and $2n + 1$ decision variables. Since b is unbounded, it can be expressed as a difference of two non-negative variables, e.g., $b = b^+ - b^-$. Since the sample median is a location parameter estimator, it is clear that it lies within the range $[\min(x), \max(x)]$, so additional constraints can be added to the problem if calculation of the range is not expensive.

The linear programming problem is efficient and has a unique solution when n is odd. When the number of observations n is even, then the problem has two alternative optimum solutions, say that, b_1^* and b_2^* are the members of two set of optimal solutions. Taking the average

$$0.5b_1^* + 0.5b_2^* = b^*$$

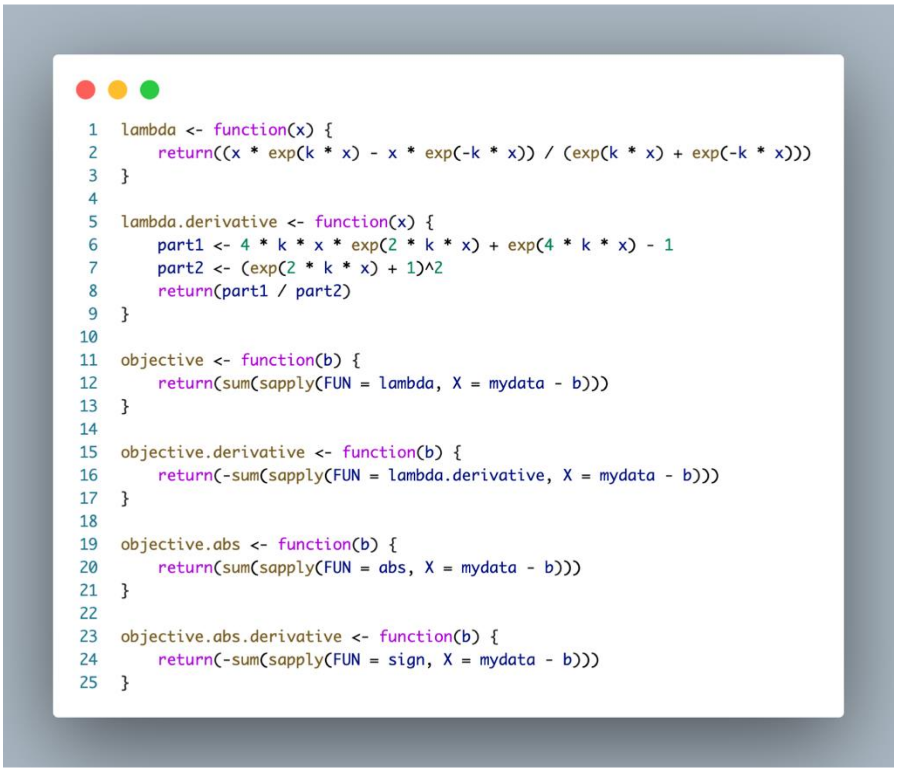
yields the sample median.

2.1. Code Snippets

In this section, median of a sample is estimated using the Formula (1) with the absolute value function itself and its approximation. Table 1 shows the compared functions and their derivatives for the estimation of the sample median. In Figure 1, R [3] implementation of the functions and their derivatives are shown. The code shown in Figure 2 performs two different optimizations for the median of the data labeled as *mydata* which has a median of 9. Figure 3 summarizes the results.

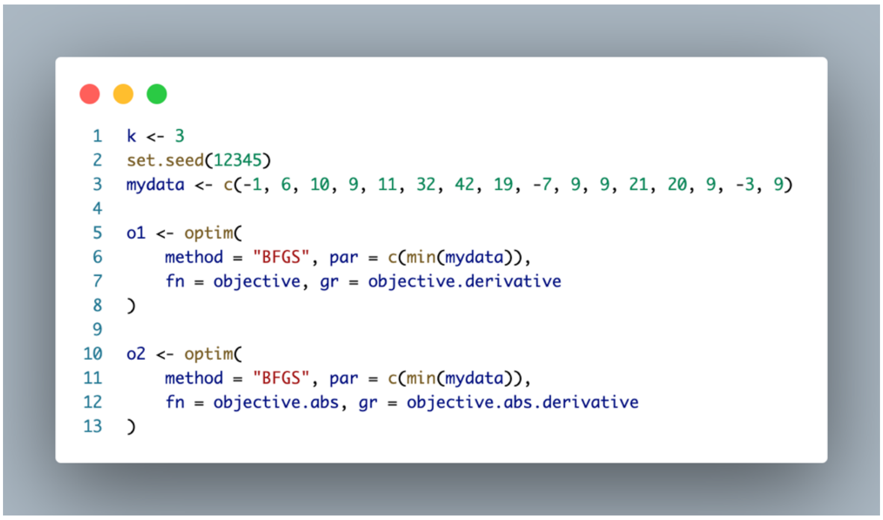
Optimization	Objective Function	Derivative	Optimizer
o1	$\lambda(x)$	$\frac{d\lambda}{dx}$	BFGS
o2	Absolute Value	Signum	BFGS

Table 1. Optimization configuration for univariate sample median



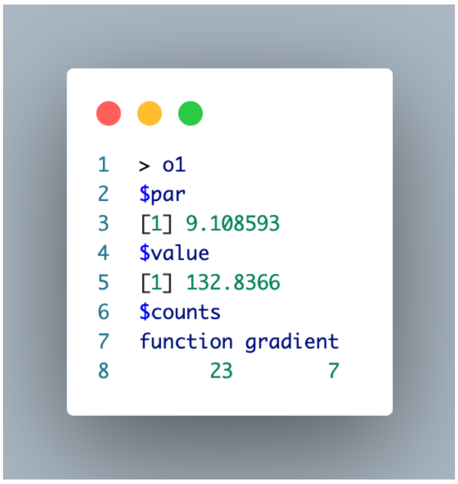
```r
lambda <- function(x) {
    return((x * exp(k * x) - x * exp(-k * x)) / (exp(k * x) + exp(-k * x)))
}

lambda.derivative <- function(x) {
    part1 <- 4 * k * x * exp(2 * k * x) + exp(4 * k * x) - 1
    part2 <- (exp(2 * k * x) + 1)^2
    return(part1 / part2)
}

objective <- function(b) {
    return(sum(sapply(FUN = lambda, X = mydata - b)))
}

objective.derivative <- function(b) {
    return(-sum(sapply(FUN = lambda.derivative, X = mydata - b)))
}

objective.abs <- function(b) {
    return(sum(sapply(FUN = abs, X = mydata - b)))
}

objective.abs.derivative <- function(b) {
    return(-sum(sapply(FUN = sign, X = mydata - b)))
}
```

Figure 1. Code Snippet 1

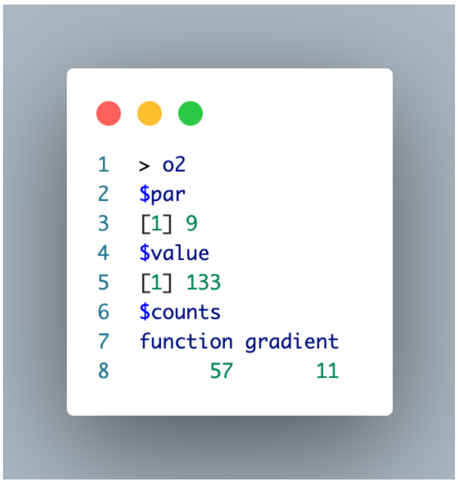


```r
k <- 3
set.seed(12345)
mydata <- c(-1, 6, 10, 9, 11, 32, 42, 19, -7, 9, 9, 21, 20, 9, -3, 9)

o1 <- optim(
    method = "BFGS", par = c(min(mydata)),
    fn = objective, gr = objective.derivative
)

o2 <- optim(
    method = "BFGS", par = c(min(mydata)),
    fn = objective.abs, gr = objective.abs.derivative
)
```

Figure 2. Code Snippet 2



```
> o1
$par
[1] 9.108593
$value
[1] 132.8366
$counts
function gradient
      23        7
```



```
> o2
$par
[1] 9
$value
[1] 133
$counts
function gradient
      57       11
```

Figure 3. R code outputs

In Figure 3, it is shown that the sample median can be estimated using a derivative based optimizer such as BFGS in R's optim() function. It is also shown that the smoothed version of absolute value function and its derivative require less function evaluations, however, the optimum solution reached is a little far from the real solution. The absolute value and signum based derivative require more function evaluations but reaches the optimum in a great precision. Note that the process of ordering 16 observations requires $n\log(n)$ in average and it is $16\text{Log}(16) \approx 44$ for this example. The gradient based approach is more efficient than the classical median() function when the number of observations are moderate and big. Interestingly, the examined estimation procedures are based on derivatives to compute sample median which is not differentiable by definition.

3. The multivariate case

Suppose $x = (\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n})$ is a collection of observations where $\boldsymbol{x_i} \in \mathbb{R}^p$, p is the dimensionality, and, n is the number of observations. The multivariate median is the center of the dataset, however, the concept "center" is not well-defined. Various multivariate median estimators are developed in the literature including the Spatial (Geometric) median [1], Oja median [9], Tukey median [10], coordinate-wise median, etc. The simplest definition is the coordinate-wise median which is based on constructing a vector of medians of each single dimension, independently. However, the vector of marginal medians does not necessarily equal to the center of data.

Geometrically, in univariate case, the sample median leaves half the data to the left and right. When $p \geq 2$, there are infinite number of directions and there is not a unique ordering of observations. That is why the sample median could not be estimated using a single pass algorithm as in the univariate case.

Let $\boldsymbol{m}$ is a p-dimensional vector of coordinates. The objective function L_b defined in Equation (1) can be re-written for p-dimensional case as

$$\min L_m = \sum_{i=1}^{n} d(\boldsymbol{x_i} - \boldsymbol{m}) \tag{2}$$

without loss of generality, where d is a distance function. By this definition, the multivariate sample median $\boldsymbol{m}$ is a new point which minimizes the total distance to all of the observations. When the distance function is chosen as the Euclidean function, the distance of a single observation is the length of the shortest line between the observation $\boldsymbol{x_i}$ and $\boldsymbol{m}$. This concept shares the same idea of *minimizing the total of distances to sample median* in univariate case as given in Equation (1) [1].

In Figure 4, a set of 2-d points on a circle with radius $r = 40$ is shown. Since the center of this circle is $(0{,}0)$, then the formula of all points in this circle is

$$x^2 + y^2 = 40^2$$

and the total distance of the points to the median is

$$\sum_{i=1}^{n} \sqrt{(x_i - m_1)^2 + (y_i - m_2)^2}$$

where m_1 and m_2 are elements of median vector, respectively. $(m_1, m_2) = (0{,}0)$ minimizes the total distances because any other points distant from the origin yields higher objective function values.

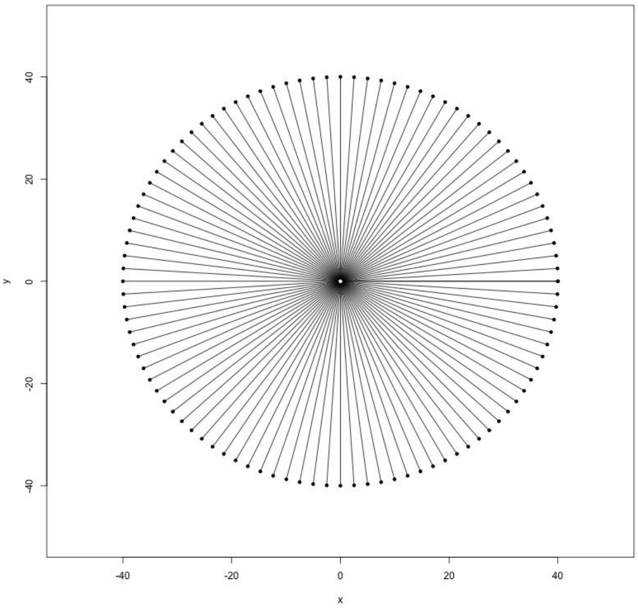


Figure 4. Observations on a circle

Similarly, assuming the number of points is infinite, the squared form of the objective function

$$\sum_{i=1}^{n} (x_i - m_1)^2 + (y_i - m_2)^2$$

has a minimum at $(m_1, m_2) = (0{,}0) = \left(\frac{\sum x_i}{n}, \frac{\sum y_i}{n}\right)$, which is equal to the sample median, but this time, the result is the sample mean just because the data is perfectly symmetric and the objective function is in squared form.

4. Least median of squares regression

Suppose the regression model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where y and x are the vectors of dependent and independent variables, respectively, β_0 and β_1 are unknown regression parameters to estimate, ϵ is the stochastic error-term with zero mean and constant variance. We have generally $n \geq p$ in regression problems so there is not a unique solution at hand. However, the ordinary least squares (OLS) estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ minimizes

$$L_{OLS} = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

so as in the way where $\bar{x}$ minimizes the univariate sample variance. In other terms, the regression estimator minimizes

$$L_{OLS} = \sum_{i=1}^{n} e_i^2$$

where $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ and $\hat{\sigma}^2 = \frac{L_{OLS}}{n-p}$, which is proportional to objective function. In short, both the sample mean and OLS estimators estimate the location parameter(s) by minimizing the variance.

The Least Median of Squares (LMS) estimator has a different loss function compared with OLS in which the summation operator is replaced by median [4]. L_{LMS} is defined as

$$L_{LMS} = \text{Median } e_i^2$$

and the LMS regression estimator $\hat{\alpha}_0$ and $\hat{\alpha}_1$ solve the problem of

$$\min_{\hat{\alpha}_0, \hat{\alpha}_1} \text{Median } (y_i - \hat{\alpha}_0 - \hat{\alpha}_1 x_i)^2.$$

Replacing the median function with its absolute value definition we yield

$$\min_{\hat{\alpha}_0, \hat{\alpha}_1, b} \sum_{i=1}^{n} |(y_i - \hat{\alpha}_0 - \hat{\alpha}_1 x_i)^2 - b|$$

where b is the median of regression residuals. Note that the parameter estimator b is not a constant but a function of residuals, i.e, $b = f(y_i - \hat{\alpha}_0 - \hat{\alpha}_1 x_i)$.

5. Least absolute deviations regression

Since the objective function of the LMS regression includes quadratic terms and it is non-linear, the estimation requires additional computational effort. Similarly, another robust regression estimator, LAD (Least Absolute Deviations) [7], minimizes the linear but non-differentiable objective function

$$L_{LAD} = \sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} \left|y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right|$$

and can be solved using many iterative optimizers, however, expressing the residuals using non-negative terms opens new rooms for effectively obtained solutions. Let

$e_1^- > 0$ if the first residual is under the regression line, whereas, $e_1^+ > 0$ if the first residual is over the regression line. If $e_1^- = e_1^+ = 0$ then the regression line exactly fits the first observation. Minimizing sum of the all deviations from the regression line implies

$$\min z = e_1^- + e_1^+ + e_2^- + e_2^+ + \cdots + e_n^- + e_n^+$$

under the constraints

$$\begin{aligned} y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1 + e_1^- - e_1^+ &= 0 \\ y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2 + e_2^- - e_2^+ &= 0 \\ &\vdots \\ y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n + e_n^- - e_n^+ &= 0 \end{aligned}$$

where

$$e_1^-, e_1^+, e_2^-, e_2^+, \dots, e_n^-, e_n^+ \geq 0$$

$\hat{\beta}_0, \hat{\beta}_1 \in \mathbb{R}$.

The linear problem has a unique solution, efficient, and easy to express. Replacing the absolute value function with non-negative deviations transforms the objective function and the constraints into linear and differentiable equations [8].

Conclusion

The sample median definition based on optimizations is similar to the definition of the sample mean and can help to better understand these two concepts. This approach can also be directly applied to estimators developed for the multivariate data. In addition to using this approach as a teaching tool, gradient-based approaches can be more efficient than order-based methods when the number of observations is large. Interestingly, the sample median can also be expressed as a linear programming problem and can be efficiently solved. Some median-based robust regression estimators can also be expressed using this approach, however, random search algorithms can be more efficient as the optimization-based approach is getting increasingly complex. Another robust regression estimator, LAD, shares the same idea of estimating the sample median in univariate case by optimizations and can be efficiently solved. Thinking the basic and the complex statistical estimators as solutions of similar optimization problems can improve understanding of basic concepts. This work does not infer any statistical results for effectiveness of the teaching approach, but presents a consistent point of view, however, a possible future work can study the impacts of different approaches.

References

[1]. H. Fritz, P. Filzmoser, & C. Croux. A comparison of algorithms for the multivariate L 1-median. Computational Statistics, 27-3 (2012): 393-410.

[2]. J.P. Paolino. "Teaching univariate measures of location-using loss functions." Teaching Statistics 40.1 (2018): 16-23.

[3]. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[4]. P. Rousseeuw. "Least median of squares regression." Journal of the American statistical association 79.388 (1984): 871-880.

[5]. A. Tali. "Minimizing the Sum of Absolute Deviations." Teaching Statistics 7.3 (1985): 88-89.

[6]. L. Yong, L. Sanyang, and Z. Shemin. "Smoothing Newton method for absolute value equations based on aggregate function." International Journal of Physical Sciences 6.23 (2011): 5399-5405.

[7]. Chen, K., Ying, Z., Zhang, H., & Zhao, L. (2008). Analysis of least absolute deviation. Biometrika, 95(1), 107-122.

[8]. Charnes, A., Cooper, W. W., & Ferguson, R. O. (1955). Optimal estimation of executive compensation by linear programming. Management science, 1(2), 138-151.

[9]. Ronkainen, Tommi, Hannu Oja, and Pekka Orponen. "Computation of the multivariate Oja median." Developments in robust statistics. Physica, Heidelberg, 2003. 344-359.

[10]. Rousseeuw, Peter J., and Ida Ruts. "Constructing the bivariate Tukey median." Statistica Sinica (1998): 827-839.