Robust method for finding the center and the scatter matrix of the cluster

Z M Shibzukhov
Instute of Mathematics and Informatics of MPGU, Moscow, Russia
Institute of Applied Mathematics and Automation KBSC RAS, Nalchick, Russia
E-mail: intellimath@mail.ru

Abstract. The problem of finding the centers and scattering matrices for a finite set of points containing outliers in a multidimensional space is considered. A new approach is considered in which instead of the arithmetic mean, differentiable mean values are used that are insensitive to outliers. An iterative reweighting scheme for searching for centers and corresponding scattering matrices for the Mahalanobis distance is considered. The examples presented in the article show the robustness property of the proposed method and algorithm with respect to a large number of outliers.

1. Introduction
The problem of finding the center and the scatter matrix for a finite set of points in a multidimensional space containing outliers is considered. To solve this problem, it is proposed to use a new approach based on minimizing differentiable estimates of the average, insensitive to outliers. Using such estimates of the mean values instead of the arithmetic mean used in the classical approach allows one to construct robust procedures for searching cluster centers and scatter matrices.

Let’s consider a finite set of points \( X = \{x_1, \ldots, x_N\} \subset \mathbb{R}^n \), which is a sample from the normal distribution with density

\[
p(x) \propto \frac{1}{\sqrt{|S|}} e^{-\frac{1}{2}(x-c)'S^{-1}(x-c)},
\]

where \( c \) is the center, \( S \) is the scatter matrix, \( |S| \) is the matrix determinant. The classical maximum likelihood method for empirically evaluating of \( c \) and \( S \) reduces to finding the minimum of the functional

\[
Q(c, S) = \ln |S| + \frac{1}{N} \sum_{k=1}^{N} D(x_k; c, S),
\]

where \( D(x) = D(x; c, S) = (x-c)'S^{-1}(x-c) \) is the square of the Mahalanobis distance. Unlike the Euclidean distance, it takes into account the difference in scales along the principal axes.

Optimal values of \( c^* \) and \( S^* \) are solutions of a system of linear equations. They can be expressed explicitly:
\[
\begin{aligned}
\begin{cases}
 c^* &= \frac{1}{N} \sum_{k=1}^{N} x_k \\
 S^* &= \frac{1}{N} \sum_{k=1}^{N} (x_k - c)'(x_k - c).
\end{cases}
\end{aligned}
\] (2)

A significant distortion of the values of \( c^* \) and \( S^* \) may appear if the empirical distribution \( \{D(x_1), \ldots, D(x_N)\} \) contains outliers. To overcome the problem, various methods and algorithms have been proposed.

The classical method aimed at solving the problem of outliers is based on replacing the square of the Mahalanobis distance function with the following function

\[
D_\varrho(x; c, S) = \varrho(\ln |S| + (x - c)'S^{-1}(x - c)),
\]

where \( \varrho(r) \) is an influence function that could suppress outliers. Usually it have to grow assymptotically slower than a linear function. The optimization task has the form:

\[
c^* = \arg \min_c \frac{1}{N} \sum_{k=1}^{N} D_\varrho(x_k; c, S).\] (3)

The problem (3) is reduced to solving systems of equations:

\[
\begin{aligned}
\begin{cases}
 c &= \frac{1}{N} \sum_{k=1}^{N} v_k x_k \\
 S &= \frac{1}{N} \sum_{k=1}^{N} v_k (x_k - c)'(x_k - c),
\end{cases}
\end{aligned}\] (4)

where

\[
v_k = \frac{\psi(D_\varrho(x_k; c, S))}{\psi(D_\varrho(x_1; c, S)) + \cdots + \psi(D_\varrho(x_N; c, S))}
\]

and \( \psi(r) = \varrho'(r) \).

For a solution to be unique, \( \psi(r) \) must be non-decreasing. But it follows from this that it is enough to make outliers at least \( \frac{1}{n+1} \)-th part of the set of points in order to break the robustness property [1]. If \( c \) is known than this method can be very robust against outliers (up to 50%). This shows that the main problem to attain high robustness is the scatter matrix.

A general family of estimators can be defined by

\[
(c, S) = \arg \min S\{D(x_1; c, S), \ldots, D(x_N; c, S)\},
\]

The condition \( |S| = 1 \) is used to rule out trivial solutions with \( |S| \rightarrow \infty \). The condition \( |S| = 1 \) means that the “shape” of the scatter is estimated.

If \( S \) is a median than there is Minimum Volume Elipsoid (MVE) estimator, and if \( S \) is a trimmed mean than there is Minimum Covariance Determinant (MCD) estimator, both proposed in [2]. A more sufficiently comprehensive review of these methods is contained in [3, 4]. In this article, we will be interested in the general case, in which \( S \) will be represented by M-means and WM-means (a sort of winzorized means). M-means and WM-means cover a very large class of well-known and practically used robust average estimates. For more detail see [5, 6].
2. Materials and methods
In this paper, we propose a new approach based on replacing the arithmetic mean in (1) with a robust differentiable estimate of the mean \( \mathbf{M}\{z_1, \ldots, z_N\} \), which will be insensitive to outliers. Such a replacement allows us to lay the foundation for the robustness of the solution of the problem at the level of mathematical formulation. This approach is an extension of a similar approach, but without the scatter matrix \( \mathbf{S} \), which was considered in [7].

The proposed approach can be considered as a generalization of the classical robust approach based on (3)–(4), when the function \( \rho \) is monotonic and invertible. The reason: the Kolmogorov mean

\[
\varrho^{-1}\left( \sum_{k=1}^{N} \rho(D_{\varrho}(x_k; c, \mathbf{S})) \right)
\]

has the same minimum as the sum

\[
\sum_{k=1}^{N} D_{\varrho}(x_k; c, \mathbf{S}).
\]

In this article we consider a robust differentiable estimates, which are constructed by two methods.

2.1. M-means
The first method is based on the approximation of the median based on the M-mean [8]

\[
\mathbf{M}_\rho\{z_1, \ldots, z_N\} = \arg \min_u \sum_{k=1}^{N} \rho(z_k - u),
\]

where \( \rho \) is a twice differentiable strictly convex function with a minimum at zero. The M-mean defined in this way has partial derivatives:

\[
\frac{\partial \mathbf{M}_\rho}{\partial z_k} = \frac{\rho''(z_k - \bar{z}_\rho)}{\rho''(z_1 - \bar{z}_\rho) + \cdots + \rho''(z_N - \bar{z}_\rho)},
\]

where \( \bar{z}_\rho = \mathbf{M}_\rho\{z_1, \ldots, z_N\} \).

For example, if we take the function \( \rho(r) = \sqrt{\varepsilon^2 + r^2} - \varepsilon \), then for sufficiently small values \( \varepsilon > 0 \), you can get an approximate and smoothed version of the median. Choosing a sufficiently small value of \( \varepsilon \), we can achieve the value

\[
\rho''(z_k - \bar{z}_\rho) = \frac{\varepsilon^2}{(\varepsilon^2 + (z_k - \bar{z}_\rho)^2)^{3/2}}
\]

will be negligible for those values of \( z_k \) that are far from the average value \( \bar{z}_\rho \).

Another example can be constructed with function \( \rho(r) = |r| - \varepsilon \ln(\varepsilon + |r|) - \varepsilon \ln \varepsilon \). So

\[
\varrho''(r) = \frac{\varepsilon}{(\varepsilon + |r|)^2}.
\]

In this case \( \varrho''(z_k - \bar{z}_\rho) \) decreases more slowly than the previous one.

In general, instead of median we can use differentiable parametric family of M-means functions based on the dissimilarity function \( \rho_\varepsilon(z - u) \) that satisfy the following requirements:

- \( \lim_{\varepsilon \to 0} \rho_\varepsilon(z - u) = |z - u| \);
- \( \lim_{\varepsilon \to 0} \rho'_\varepsilon(z - u) = \text{sign}(z - u) \);
- \( \lim_{\varepsilon \to 0} \rho''_\varepsilon(z - u) = 0 \).
2.2. WM-mixed

The second method is based on the use of the approximation of the winzorized-like mean

$$WM_{\rho,\alpha}\{z_1, \ldots, z_N\} = \frac{1}{N} \sum_{k=1}^{N} \min\{z_k, \bar{z}_{\rho,\alpha}\},$$

where $\rho(r)$ is the function for which $M_{\rho}$ acts as a differentiable approximation of the median,

$$\rho_{\alpha}(r) = \begin{cases} 
\alpha \rho(r), & \text{if } r > 0 \\
\frac{1}{2}(\alpha \rho(0^+) + (1 - \alpha)\rho(0^+)), & \text{if } r = 0 \\
(1 - \alpha)\rho(r), & \text{if } r < 0 
\end{cases}$$

is a function for which $M_{\rho_{\alpha}}\{z_1, \ldots, z_N\}$ acts as a differentiable approximation of $\alpha$-quantile.

Partial derivatives are of the form:

$$\frac{\partial WM_{\rho_{\alpha}}}{\partial z_k} = \begin{cases} 
\frac{1}{N} \frac{m \partial M_{\rho_{\alpha}}}{\partial z_k}, & \text{if } z_k \leq \bar{z}_{\rho_{\alpha}} \\
\frac{m \partial M_{\rho_{\alpha}}}{N \partial z_k}, & \text{if } z_k > \bar{z}_{\rho_{\alpha}} 
\end{cases}$$

where $m$ is the number of values $z_k > \bar{z}_{\rho_{\alpha}}$.

For example, there is also a smooth variant of $\min\{z, u\}$:

$$\sigma(z, u) = \frac{1}{2}(z + u - \rho_{\alpha}(z - u)),$$

where $\rho$ is a function, which is used for construction $M$-average as replacement of median (for example, $\rho_{\text{sqrt},\epsilon}, \rho_{\ln,\epsilon}, \rho_{\arctan,\epsilon}$). So

$$WM_{\rho,\alpha}\{z_1, \ldots, z_N\} = \frac{1}{N} \sum_{k=1}^{N} \sigma(z_k, \bar{z}_{\rho_{\alpha}})$$

In this case

$$\frac{\partial WM_{\rho_{\alpha}}}{\partial z_k} = \frac{1}{2N} (1 - \rho'(z_k - \bar{z}_{\rho_{\alpha}})) + \frac{1}{2} \frac{\partial M_{\rho_{\alpha}}}{\partial z_k}.$$
2.3. The principle of minimizing robust estimates

Thus, in terms of outliers, one can search for $c^*$ and $S^*$, minimizing the functional

$$Q(c, S) = M\{z_1(c, S), \ldots, z_N(c, S)\},$$

where $M$ is one of the above differentiated estimates of the average value,

$$z_k = z_k(c, S) = \ln |S| + D(x_k; c, S).$$

By virtue of differentiability,

$$\begin{align*}
    \mathbf{v} &= \nabla M\{z_1, \ldots, z_N\} \\
    \mathbf{c} &= \sum_{k=1}^{N} v_k x_k \\
    \mathbf{S} &= \sum_{k=1}^{N} v_k (x_k - \mathbf{c})'(x_k - \mathbf{c}).
\end{align*}$$

The weight vector $\mathbf{v}$ for $\mathbf{c} = c^*$ and $\mathbf{S} = S^*$ can also be used as an estimate of the significance of points. Since $v_1 + \cdots + v_N = 1$, the outliers will correspond to the points with the lowest values of the weights.

Stability with respect to outliers is achieved due to the fact that the weights of the points corresponding to outliers are significantly less than the weights of the points that are not outliers. One also can to see that in classical approach $v_k$ is calculated as function of the distance value. In contrast in our approach $v_k$ is function of the difference between the distance and the robust mean value $c$ of the distances. As a result the weights of the points, which are nearer to the center $c$, become significantly larger than the weights of the outliers. This is decrease the influence of the outliers since their weights become relatively smaller. Such properties are a natural consequence of the robustness of average estimates. Differentiability provides the possibility of using iterative gradient procedures to find a solution.

This approach is similar to that used in [10].

2.4. The algorithm

To search for $c^*$ and $S$, we apply an iterative scheme that corresponds to the Jacobi method for solving the system of nonlinear equations (9):

(i) At the beginning,

$$\begin{align*}
    \mathbf{c}_0 &= \frac{1}{N} \sum_{k=1}^{N} \mathbf{x}_k \\
    \mathbf{S}_0 &= \mathbf{I},
\end{align*}$$

where $\mathbf{I}$ is the identity matrix $n \times n$.

(ii) At the $t$-th step, two equations are successively solved:

(a) First, the following equation is solved for finding $\mathbf{c}_{t+1}$:

$$\mathbf{c} = \sum_{k=1}^{N} \frac{\partial M\{z_1, \ldots, z_N\}}{\partial z_k} \mathbf{x}_k, \quad (10)$$

where $z_k = D(x_k; c, \mathbf{S}_t)$. 


(b) Second, the following equation is solved to find $S_{t+1}$:

$$S = \sum_{k=1}^{N} \frac{\partial M(z_1, \ldots, z_N)}{\partial z_k} (x_k - c_{t+1})' (x_k - c_{t+1}),$$

where $z_k = D(x_k; c_{t+1}, S)$.

(iii) Step 2 is repeated until $t < T$ (maximum number of iterations) or the sequence $\{Q(c_t, S_t)\}$ will not concentrate around its thickening point.

In particular cases additional condition $|S| = 1$ can be added in order to prevent singularity of the correlation matrix. The scale factor $\sigma = |S|$ can be estimated using $S$-estimator [9].

First equation has a form

$$c_{t+1} = (1 - h)c_t + hF(c_t).$$

In order to find its solution we used the following simple iterative procedure:

$$S_{t+1} = (1 - h)S_t + hF(S_t).$$

In both cases $0 < h < 1$ – a step value.

For numerical calculation the implementation of this algorithm in the package mlgrad (https://bitbucket.org/intellimath/mlgrad) is used.

3. Results and discussions

First, consider examples with small real data sets that are commonly used to demonstrate the robustness of robust center search methods and the scatter matrix. They use the square of the Mahalanobis distance $D(x; c, S) = (x - c)' S^{-1} (x - c)$. In the classic distance, the center and matrix are used, calculated by (2). In the robust distance, the center and matrix are used, calculated according to the above iterative reweighted scheme.

As a robust differentiable estimate of the mean, a winzorized mean (6) is used, which uses the $M$-mean $M_\rho$ with the function $\rho_\alpha$ of the form (7). Moreover, $\rho(r) = \sqrt{r^2 + r^2 - \varepsilon}$. The value of the parameter $\alpha$, as a rule, is associated with the value of the proportion of objects that are not outliers.

**Example 1.**

We consider the stackloss dataset (projection on the attributes watertemp and stackloss). The center and correlation matrix are searched. Figure 1 shows the found centers and spanning ellipses. The first ellipse covers all points. The second ellipse corresponds to the robust scatter matrix. It covers all points except the outliers. Its largest semiaxis passing through the center can also be considered as a linear regression line that models these non-outlier points.

A scatter chart is also provided with which outliers can be identified. The non-outlier points are concentrated along the line near bottom of the figure. The coordinates of the points in this scatter chart are classic and robust distances. The classic distance is calculated using the scatter matrix (2). The robust distance is calculated using the scatter matrix, which is a solution to the system of equations (9). Parameters $\alpha = 0.8$ and $\varepsilon = 1.0$. 
Example 2.
We consider the starsCYR dataset. The projection on the attributes \( \log.Te \) and \( \log.light \) is considered. The center and correlation matrix are searched. Figure 2 shows the found centers and spanning ellipses. The first ellipse also covers all points. The second ellipse, corresponding to the robust correlation matrix, covers all points except the outliers. It covers all points except the outliers too. Its largest semiaxis passing through the center can also be considered as a linear regression line that models these non-outlier points.

A scatter chart is also provided with which outliers can be identified. The coordinates of the points are classic and robust value of the distances. The classic distance is calculated using the scatter matrix (2). The robust distance is calculated using the scatter matrix, which is a solution to the system of equations (9). Parameters \( \alpha = 0.8 \) and \( \varepsilon = 1.0 \).

Example 3.
A stackloss dataset with a full set of attributes (4 attributes) is considered. In the 4-dimensional space of attribute values, the center and the scatter matrix are searched. Figure 3 shows the
results. The first two figures show the empirical distribution of the values of the classical and robust distances. According to them, 1 and 4 outliers can be separated, respectively. The third figure shows a scattering diagram for two parameters: the classical and robust distances to their centers. On it, you can confidently identify all 5 emissions. Here the parameters $\alpha = 0.8$ and $\varepsilon = 1.0$.

**Figure 3.** Comparison of classic and robust methods (stackloss, 4 attributes).

**Example 4.**

A wine dataset with a full set of attributes (13 attributes) is considered. There are scatter plots for each class in the figure 4. For each class a classical and robust center and scatter matrix are estimated. First axis corresponds to the distance with the classical center and the scatter matrix. Second axis corresponds to the distance with the robust center and the scatter matrix. It is easy to separate outliers from non-outliers with the help of these scatter plots. The robust center and the scatter matrix are estimated by our algorithm with the parameters $\alpha = 0.8$ and $\varepsilon = 1.0$.

**Figure 4.** Comparison of classical and robust methods (wine).
Example 5.
A boston dataset is being considered. Each object is also described by multidimensional vector. There are two scatter plots in figure 5. First axis in the scatter plots corresponds to the distance with the classical center and the scatter matrix. Second axis corresponds to the distance with the robust center and the scatter matrix. Second scatter plot is magnified version of the first one. It used to explain the ability to discover a second group of points – a possible second cluster. This is allow to conclude that there are at least two clusters in boston dataset. It is easy to identify outliers from it. Parameters $\alpha = 0.8$ and $\varepsilon = 1.0$, as in previous cases.

![Figure 5. Comparison of classical and robust methods (boston).](image)

Example 6.
This example is artificial. It demonstrates the stability of the method with respect to a relatively large volume of outliers with different data dimensions ($n = 2, 10, 20$). Normal samples ($N = 1000$) are generated from the multidimensional normal distributions and the outlier samples ($M = 800$) are generated from another normal distribution. First and second samples are not intersected. There is sufficient margin between them. There are 44% of outliers. Figure 6 shows a scatterplot. First axis in the scatter plots corresponds to the distance with the classical center and the scatter matrix. Second axis corresponds to the distance with the robust center and the scatter matrix. Normal points are in the bottom of the plots alone the line. Outliers are clearly separated from normal points. Presented here scatter plots clearly explain resistance of our algorithm to large number of outliers within relatively large dimensions.

![Figure 6. Comparison of classical and robust methods (multidimensional normal distribution).](image)
Example 7.
This example is also artificial. It differs from the previous method of generating emissions. In this case, \(n\) (equals to the number of coordinates of the points) separate samples are generated from the multidimensional normal distribution, with the centers located on the axes (i.e. the coordinates of the center, except one, are equal to zero). Each additional sample contain \(800/n\) points. So there are 800 outliers. A such method produces a dataset with the cells (outliers) [11]. There are scatter plots in figure 7. shows scatter charts that also shows the method’s resistance to large outliers. However this case differs from previous example. We need to apply 2-dimensional \textit{k-means} algorithm in order to separate regular points from the outliers.

![Figure 7. Comparison of classical and robust methods (multidimensional normal distribution).](image)

4. Conclusion
A new approach is proposed for the empirical estimation of the center and scattering matrix for a finite set of points containing outliers. The peculiarity of the approach is that instead of the arithmetic mean, differentiable M-means are used to construct a smoothed version of the median and quantile, and WM-means are used to construct a smoothed version of the winzorized average, in which the value of the smoothed quantile estimate is used as the threshold value. The parameter of the smoothed \(\alpha\)-quantile level corresponds to the proportion of points that do not contain outliers. The examples considered are based on real data, which are usually used to demonstrate the reliability of methods and algorithms for finding the center and scatter matrix, as well as artificial examples that convincingly show the stability of the proposed method and algorithm to large volumes of outliers.

Acknowledgments
This work is supported by the grant of RFBR 18-01-00050.

References
[1] Maronna R A 1976 Robust M-Estimators of Multivariate Location and Scatter Ann. Statist. 4 51–67
[2] Rousseeuw P J 1985 Multivariate estimation with high breakdown point Mathematical Statistics and Applications B, ed W Grossmann, G Pflug, I Vincze and W Wertz (Reidel, Dordrecht) pp 283–297
[3] Rousseeuw P and Hubert M 2013 High-breakdown estimators of multivariate location and scatter Robustness and Complex Data Structures ed C Becker, R Fried and S Kuhnt pp 49–66
[4] Maronna R A and Yohai V J 2015 Robust and efficient estimation of multivariate scatter and location Preprint arxiv:1504.03389
[5] Huber P J 1981 Robust Statistics (John Wiley & Sons, New York)
[6] Beliakov G Sola H and Calvo T 2016 A Practical Guide to Averaging Functions (Springer)
[7] Shibzukhov Z M and Kazakov M A 2019 Clustering based on the principle of finding centers and robust averaging functions of aggregation Journal of Physics: Conference Series 1368 052010
[8] Shibzukhov Z M 2017 On the principle of empirical risk minimization based on averaging aggregation functions *Doklady Mathematics* 96(2) pp 494–497

[9] Davies P L 1987 Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices *Ann. Statist.* 15 pp 1269–1292

[10] Shibzukhov Z M, Kazakov M A and Dimitrichenko D P 2019 Robust training of neural network via minimizing robust estimates of the average of loss functions *Journal of Physics: Conference Series* 1203 012074

[11] Agostinelli C, Leung A, Yohai V J, Zamar R H 2014 Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination *Preprint* arxiv:1406.6031