Comparison of detection of outliers using classic and robust Mahalanobis distance

Concepts and examples in R with MCD and MVE estimators

Final BSc Project

João Carlos Pereira Fernandes
92635, LMAC

Advisor: Isabel Maria Alves Rodrigues

2020/2021

Contents

1 Introduction to robust statistics

1.1 Outliers in multivariate data

Consider here a p-variate n-data set (data set with n observations and p different variables) $X = \{x_1, x_2, ..., x_n\}$.

Definition 1.1 (Outlier)

For a certain $1 \leqslant i \leqslant n$, we say x_i is an **outlier** if its value deviates from the fit suggested by the majority of the observations.

The definition of outlier is intuitive but the flagging of outliers depends on the considered estimator. The event of an actual outlier not being flagged is called **masking** and when a regular observation is taken as an outlier it's called **swamping**.

When dealing with multivariate data, an outlier cannot be determined by looking at each variable individually. Multivariate outliers can only be detected by correctly estimating the covariance structure.

Definition 1.2 (Mahalanobis Distance)

Given a p-dimensional center vector μ and a covariance matrix Σ of size p, the **Mahalanobis distance** of $x \in \mathbb{R}^p$ is given by

$$\mathrm{MD}(x, \mu, \Sigma) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

The Mahalanobis distance is a useful tool for multivariate outlier detection. When μ is equal to the sample mean and Σ is equal to the identity matrix, the Mahalanobis distance becomes the Euclidean distance from x to μ. This distance is more reliable because it takes into consideration the covariance between the variables.

Proposition 1.1) If the data set X follows a p-normal distribution, that is $X \sim N_p(\mu_X, \Sigma_X)$, then the squared Mahalanobis distance follows a chi-squared distribution with p degrees of freedom, that is $\mathrm{MD}^2(x, \mu_X, \Sigma_X) \sim \chi^2_p$

The proof for this proposition can be found in [8] and it comes from simple algebraic manipulation and transformation of variables.

Definition 1.3 (Tolerance ellipsoid)

The **tolerance ellipsoid** is one of the most common tools for multivariate outlier detection. Given a location parameter μ and a scatter parameter Σ, it is defined as

$$\{x \in \mathbb{R}^p : \mathrm{MD}(x, \mu, \Sigma) \leqslant \sqrt{\chi^2_{p,0.975}}\}$$

where $\chi^2_{p,0.975}$ gives the 97.5% quantile of the chi-squared distribution with p degrees of freedom. The sample **classical tolerance ellipsoid** is obtained by considering μ as the mean vector of all observations and Σ the sample covariance matrix of all observations.

As a result from Proposition 1.1), as the number of observations n grows to infinity, about 97.5% of them will be contained by the ellipsoid. Any observation not contained within the ellipsoid gets flagged as an outlier.

As it will be seen ahead, the main difference in the classical and robust tolerance ellipsoids used for detection of outliers are the estimates of μ, Σ and the number of observations considered for their calculation.

1.2 Estimators and properties

Definition 1.4 (Estimation, Estimator)

Estimation is a procedure used to calculate the value of some property of a population from a pool of observations drawn from it. An **estimator** is any function, algorithm or rule used to calculate an estimate.

Definition 1.5 (Robustness, Efficiency)

The notion of **robustness** of an estimator is a measurement of how they are influenced by outliers: an estimator is more robust if they are less influenced by outliers. **Efficiency** measures the precision of an estimator with uncontaminated data: an estimator is more efficient if the gives the most plausible results for a general uncontaminated data set.

In data analysis, it is very common to find data sets which have outliers that can result in distorted estimates. Robust statistics aims to find estimators that combine high robustness with high efficiency.

For the purposes of this work, we assume our observations x_i are independent and identically distributed with a distribution function F. The most common model is the multivariate normal distribution, that is $x_i \sim N_p(\mu, \Sigma)$. In this context, μ is called the **location** parameter and Σ is the **scatter** parameter. The estimators discussed in this paper aim to determine the most accurate estimate of the location and scatter parameters and identify outliers accordingly.

Definition 1.6 (Affine equivariance)

We say that a location estimator T_μ and a scatter estimator T_Σ are **affine equivariant** if they check the following condition:

$$T_\mu(\{Ax_1 + b, ...Ax_n + b\}) = A\,T_\mu(\{x_1, ...x_n\}) + b$$
$$T_\Sigma(\{Ax_1 + b, ..., Ax_n + b\}) = A\,T_\Sigma(\{x_1, ..., x_n\})\,A^T$$

for any invertible matrix A and vector b (with compatible sizes with the data set).

Affine equivariance means that the estimators transform well under any reparametrization of the space of the observations x_i, that is, data can be rescaled, rotated or translated without affecting the detection of outliers.

1.3 Measures of robustness

Definition 1.7 (Breakdown value - BV)

Given a n-data set X_n, the **breakdown value** is the smallest fraction of observations that need to be replaced by arbitrary values before it generates unreasonable estimates. For a multivariate **location** estimator T_μ it is defined as:

$$\varepsilon_n^*(T_\mu, X_n) = \frac{1}{n} \min\{m : \sup \,||\, T_\mu(X_n) - T_\mu(X_{n,m}) \,|| = +\infty\}$$

where the supremum is taken over all sets $X_{n,m}$ obtained by replacing m observations of X_n by arbitrary values. For a multivariate **scatter** estimator T_Σ it is defined as:

$$\varepsilon_n^*(T_\Sigma, X_n) = \frac{1}{n} \min\{m : \sup \max_i |\, \log(\lambda_i(T_\Sigma(X_n))) - \log(\lambda_i(T_\Sigma(X_{n,m}))) \,| = +\infty\}$$

where $X_{n,m}$ has the same meaning and $\lambda_i(T_\Sigma(X_n)), i = 1, ..., p$ represent the eigenvalues of T_Σ. Note that a scatter estimator generates a positive definite matrix and therefore all of its eigenvalues are positive. A scatter estimator is considered to be broken when an eigenvalue is arbitrarily close to zero (implosion) or to infinity (explosion).

Example 1.1 (Breakdown value of the median)

Let $X = \{x_1, ..., x_n\}$ and $T(X) = med(X)$.

If we replace $\left\lfloor \frac{n-1}{2} \right\rfloor$ observations by any value, $T(X^*) \in [x_{(1)}, x_{(n)}]$. However, replacing $\left\lfloor \frac{n+1}{2} \right\rfloor$ observations by $a > x_{(n)}$ will yield $T(X^*) = a \notin [x_{(1)}, x_{(n)}]$. Since we can choose a arbitrarily large, $T(X^*)$ can't be bounded.

Therefore, the BV of the median is $\varepsilon_n^*(T, X) = \frac{1}{n} \left\lfloor \frac{n+1}{2} \right\rfloor \approx 50\%$

Definition 1.8 (Sensitivity curve - SC)

Given an estimator T and $n-1$ fixed observations of a data set $X_{n-1} = \{x_1, ..., x_{n-1}\}$, the **sensitivity curve** measures the effect when adding a single observation x to the data set. It is defined as:

$$SC(x, T_n, X_{n-1}) = \frac{T_n(x_1, ..., x_{n-1}, x) - T_{n-1}(x_1, ..., x_{n-1})}{\frac{1}{n}}$$

Note that T_k means that we are applying the estimator T to a k-data set.

Definition 1.9 (Influence function - IF)

The **influence function** is the asymptotic equivalent of the SC. It is computed for an estimator T at a certain distribution F and does not depend on the data set. Given x and the distribution $F_{\varepsilon,x} = (1-\varepsilon)F + \varepsilon\delta_x$, for $\varepsilon > 0$, where δ_x is the distribution that gives all its mass to x, we define the influence function as:

$$IF(x,T,F) = \lim_{\varepsilon \to 0} \frac{T(F_{\varepsilon,x}) - T(F)}{\varepsilon} = \frac{\partial}{\partial \varepsilon} T(F_{\varepsilon,x}) \Big|_{\varepsilon=0}$$

Definition 1.10 (Gross-error sensitivity - GES)

The **gross-error sensitivity** is defined as the supremum of the absolute value of the IF over all points where it's defined.

$$\gamma^*(T,F) = \sup_x |IF(x,T,F)|$$

Definition 1.11 (Asymptotic normality, Asymptotic variance - AV)

An estimator is said to be **asymptotically normal** if the sampling distribution follows a normal distribution as the sample size n grows to infinity. Formally, it is asymptotically normal for an estimate μ if $\sqrt{n}(\mu - T_n)$ converges in distribution to $N(0, \sigma_{T_n})$. For an asymptotically normal estimator T at a distribution F, we define the **asymptotic variance** as:

$$V(T,F) = \int IF(x,T,F)^2 \, dF(x)$$

Definition 1.12 (Asymptotic efficiency - AE)

The **asymptotic efficiency** of an estimator T at a distribution F is defined as

$$\text{eff}\,(T,F) = \frac{1}{V(T,F)\, I(F)}$$

where $I(F)$ is called the **Fisher information** of the model and is defined as $I(F) = \int(-f'(x)/f(x))^2 dF(x)$

A "good" estimator is, in general, one that combines:

- High breakdown value
- Bounded influence function
- Small gross-error sensitivity
- Small asymptotic variance
- High asymptotic efficiency

In this paper, there will be studied two different highly robust estimators of multivariate location and scatter: the Minimum Covariance Determinant (MCD) estimator and the Minimum Volume Ellipsoid (MVE) estimator.

2 MCD estimator

The Minimum Covariance Determinant estimator, introduced in 1984, is a highly robust estimator of multivariate location and scatter.

Definition 2.1 (MCD estimator)

Given a p-variate n-data set, for a fixed h such that $\frac{n+p+1}{2} \leqslant h \leqslant n$, the MCD estimations of location and scatter are $\hat{\mu}$ and $\hat{\Sigma}$, respectively, where:

- $\hat{\mu}$ is the mean of the h observations for which the determinant of the sample covariance matrix is minimal;
- $\hat{\Sigma}$ is that covariance matrix with minimal determinant, multiplied by a consistency factor.

h can be seen as the number of points accounted for the choice of μ and Σ, and then we can consider $\alpha = \frac{h}{n}$ as the proportion of observations used.

2.1 Properties

Proposition 2.1) The MCD estimator is affine equivariant.

Proof: Let $X = \{x_1, ..., x_n\}$ be a p-variate n-data set and $X_{(h)}$ any h-subset of X. Furthermore, denote by $AX + b = \{Ax_1 + b, ..., Ax_n + b\}$ the data set obtained by applying the function $x \mapsto Ax + b$ to every observation of X, for any $p \times p$ non-singular matrix A and $p \times 1$ vector b. By the properties of the covariance matrix, denoting by $\Sigma(X)$ the covariance matrix of X, we get:

$$\begin{aligned}
\Sigma(AX + b) &= E[(AX + b - E[AX + b])\,(AX + b - E[AX + b])^T] = \\
&= E[(AX) + b - E[AX] - b)\,(AX + b - E[AX] - b)^T] \\
&= E[(A(X - E[X]))\,(A(X - E[X]))^T] = \\
&= E[A\,(X - E[X])\,(X - E[X])^T\,A^T] = \\
&= A\;E[(X - E[X])\,(X - E[X])^T]\;A^T = \\
&= A\,\Sigma(X)\,A^T
\end{aligned}$$

Since for any matrices A and B (with compatible sizes for multiplication) we have $|AB| = |A|\,|B|$ and $|A| = \left|A^T\right|$, we can see that $|\Sigma(AX + b)| = \left|A\Sigma(X)A^T\right| = |A|^2\,|\Sigma(X)|$ and therefore, if the h-subset $\{x_1, ..., x_h\}$ minimizes $\left|\Sigma(X_{(h)})\right|$, then the subset $\{Ax_1 + b, ..., Ax_h + b\}$ minimizes $\left|\Sigma(AX_{(h)} + b)\right|$.

Note that this proof does not depend on h, which means the scatter estimator is affine equivariant.

As for the proof of the affine equivariance of the location estimator, notice that $E(AX + b) = AE(X) + b$, by linearity of the mean, and then the proof becomes analogous to the scatter estimator.

Proposition 2.2) The MCD estimator has a bounded IF.

The proof of this proposition is lengthy and can be found in full detail in [5].

The main idea is to first prove that the MCD solution at the contaminated distribution $F_{\varepsilon,x}$ is still determined by an ellipsoid. We can consider a distribution function with zero mean vector and identity matrix covariance because the MCD estimator is affine equivariant. Afterwards, conclude that the IF of the scatter matrix part is given by:

$$IF(x, \Sigma, F) = \frac{-1}{2c}\,x\,x^T\,\mathbf{1}_{\{||x||^2 \leqslant q_\alpha\}} + w(||x||)\,I_p$$

for constants c and q_α (α denotes the portion of observations used for the computation of the MCD estimator, $0 < \alpha < 1$), a real valued bounded function w and where I_p denotes the identity matrix of size p. This makes it easy to see that the IF is bounded. As for the location estimator, its IF can be written as:

$$IF(x, \mu, F) = \left(\frac{-2}{\alpha}\int_{\{||z|| \leqslant q_\alpha\}} z\,z^T g'(||z||)\,dz\right)^{-1}\frac{x}{\alpha}\mathbf{1}_{\{||x||^2 \leqslant q_\alpha\}}$$

for a certain function g with strictly negative derivative g'. The IF is then zero outside of an ellipsoid, and bounded inside that ellipsoid, making the IF bounded in all of its domain.

Definition 2.2 (General position)

A p-variate data set is said to be in **general position** if at most p observations lie in a $(p-1)$-dimensional hyperplane.

When we are sampling from a continuous distribution, as for example the multivariate normal distribution, almost surely the data set will be in general position.

Proposition 2.3) If the data is in general position, the MCD estimators of location and scatter will have a BV equal to

$$\varepsilon_n^*(T_\mu, X) = \varepsilon_n^*(T_\Sigma, X) = \frac{\min(n-h+1,\, h-p)}{n}$$

This can even be generalized when the data set is not in general position. Denoting by k_X the maximum number of observations on any hyperplane of $\mathbb{R}^p$, if $k_X < h$ it suffices to swap p for k_X.

Note that in the previous proposition, and using $\alpha = \dfrac{h}{n}$, we have that

$$\lim_{n\to\infty} \varepsilon_n^*(T_\mu, X) = \lim_{n\to\infty} \varepsilon_n^*(T_\Sigma, X) = \min(\alpha,\, 1-\alpha)$$

and thus maximum BV is achieved when $\alpha = 0.5$, that is, $h = \dfrac{n+p+1}{2}$.

2.2 FAST-MCD and DetMCD

The exact algorithm to calculate the MCD estimations of location and scatter implies looking at all the h-subsets of the n-data set, computing the mean and covariance matrix for each one and retaining the subset which leads to smallest covariance matrix determinant. This becomes very computationally expensive for very large or higher dimensional data sets. Thus, there are two main alternative algorithms to compute the MCD estimations: **FAST-MCD** and **DetMCD**.

FAST-MCD algorithm

If n is small, say $n \leqslant 500$:

1. Draw a random subset of size $p+1$ and compute its mean and covariance matrix.
2. Apply two **C-steps**:
   - Compute robust distances based on the most recent mean and covariance estimate.
   - Take the h observations with smallest robust distance.
   - Compute mean and covariance matrix of this h-subset.
3. Retain the 10 h-subsets with smallest covariance determinant.
4. Apply C-steps on these subsets until convergence.
5. Retain the h-subset with smallest covariance determinant.

If n is large, say $n > 500$:

1. Draw $m \leqslant 5$ disjoint subsets of size n_{sub} (note that they should be sufficiently large but don't necessarily need to amount to the entire data set).
2. In each subset, repeat 100 times:
   - Construct an initial subset H_1 of size $h_{sub} = \dfrac{n_{sub}}{n}\, h$
   - Apply two C-steps to H_1 using n_{sub} and h_{sub}
   - Retain the 10 best solutions $(\hat{\mu}_{sub}, \hat{\Sigma}_{sub})$
3. Merge the m subsets into a subset of size n_{merge} and for each of $10m$ solutions:
   - Apply two C-steps using n_{merge} and $h_{merge} = \dfrac{n_{merge}}{n}\, h$
   - Retain the 10 best solutions $(\hat{\mu}_{merge}, \hat{\Sigma}_{merge})$
4. In the full dataset, apply C-steps (preferably until convergence) using n and h to the $k \leqslant 10$ best solutions and the get the final estimation $(\hat{\mu}, \hat{\Sigma})$ which minimizes $|\Sigma|$

Proposition 2.4) C-steps decrease (or equal, when it reaches convergence) the covariance matrix determinant.

Proof: Consider a p-variate n-data set $X = \{x_1, ..., x_n\}$ and H_1 a random h-subset of X. Let $\hat{\mu_1}$ denote the mean vector of H_1 and $\hat{\Sigma_1}$ the sample covariance matrix of H_1. Computing the robust distances defined as:

$$d_1(x_i) = \sqrt{(x_i - \hat{\mu}_1)^T \, \hat{\Sigma}_1^{-1} \, (x_i - \hat{\mu}_1)}, \; i = 1, ..., n$$

we can obtain $H_2 = \{x_{(1)}, ..., x_{(h)}\}$ as the h-subset of X such that the observations in H_2 have the smallest robust distance d_1, that is $d_1(x_{(1)}) \leqslant d_1(x_{(2)}) \leqslant ... \leqslant d_1(x_{(n)})$. Denoting by $\hat{\mu_2}$ the mean vector of H_2 and $\hat{\Sigma_2}$ the sample covariance matrix of H_2, we want to prove that $|\hat{\Sigma_2}| \leqslant |\hat{\Sigma_1}|$. Consider the following equality:

$$\begin{aligned}
\frac{1}{hp} \sum_{x \in H_2} (d_2(x))^2 &= \frac{1}{hp} \operatorname{Tr} \sum_{x \in H_2} (x - \hat{\mu}_2) \hat{\Sigma}_2^{-1} (x - \hat{\mu}_2)^T = \\
&= \frac{1}{p} \operatorname{Tr} \, \hat{\Sigma}_2^{-1} \sum_{x \in H_2} \frac{1}{h} (x - \hat{\mu}_2)(x - \hat{\mu}_2)^T = \\
&= \frac{1}{p} \operatorname{Tr} \, \hat{\Sigma}_2^{-1} \, \hat{\Sigma}_2 = \frac{1}{p} \operatorname{Tr}(I_p) = 1
\end{aligned}$$

Now, define $\lambda := \dfrac{1}{hp} \displaystyle\sum_{x \in H_2} (d_1(x))^2$ and because H_2 was chosen to have the observations with smallest distance d_1, we have:

$$\lambda = \frac{1}{hp} \sum_{x \in H_2} (d_1(x))^2 \leqslant \frac{1}{hp} \sum_{x \in H_1} (d_1(x))^2 = 1$$

Using both results yields:

$$\frac{1}{hp} \sum_{x \in H_2} (d_{(\mu_1, \lambda \Sigma_1)}(x))^2 = \frac{1}{hp} \sum_{x \in H_2} (x - \hat{\mu}_1) \frac{1}{\lambda} \hat{\Sigma}_1^{-1} (x - \hat{\mu}_1)^T = \frac{1}{\lambda hp} \sum_{x \in H_2} (d_1(x))^2 = \frac{\lambda}{\lambda} = 1$$

Grübel (see [9]) has proven that $(\hat{\mu}_2, \hat{\Sigma}_2)$ is the unique minimizer of $|\hat{\Sigma}|$ over all $(\hat{\mu}, \hat{\Sigma})$ that satisfy $\dfrac{1}{hp} \displaystyle\sum_{x \in H_2} (d_{(\mu, \Sigma)}(x))^2 = 1$. From this it follows that $|\hat{\Sigma}_2| \leqslant |\lambda \hat{\Sigma}_1|$ and because we have seen that $\lambda \leqslant 1$, it follows immediately that $|\lambda \hat{\Sigma}_1| \leqslant |\hat{\Sigma}_1|$. Therefore:

$$|\hat{\Sigma}_2| \leqslant |\lambda \hat{\Sigma}_1| \leqslant |\hat{\Sigma}_1|$$

So, we get $|\hat{\Sigma}_2| \leqslant |\hat{\Sigma}_1|$. Notice that by Grübel's proof and the definition of λ, we have $|\hat{\Sigma}_2| = |\hat{\Sigma}_1|$ if and only if $(\hat{\mu}_1, \hat{\Sigma}_1) = (\hat{\mu}_2, \hat{\Sigma}_2)$, which implies convergence.

DetMCD algorithm

Here we take the data set X as the $n \times p$ matrix where each row represents a different observation and each column a different variable.

1. Draw six initial estimates of location and scatter $(m_k, S_k), \; k = 1, ..., 6$
2. For each estimate (m_k, S_k):
   - Compute the matrix E_k of the eigenvectors of S_k and let $B_k = X E_k$
   - Estimate the covariance matrix by $\Sigma_k = E_k \, L \, E_k^T$, where $L = \operatorname{diag}(Q_n(B_{k_1})^2, ..., Q_n(B_{k_p})^2)$ and Q_n represents the scale estimator defined as $Q_n(X) = 2.219\{|x_i - x_j| : \; i > j\}_{(k)}$ with $k = \binom{\lceil \frac{n}{2} \rceil + 1}{2}$
   - Estimate the location $\hat{\mu_k} = \hat{\Sigma}_k^{\frac{1}{2}} \; \operatorname{med}(X \, \hat{\Sigma}_k^{-\frac{1}{2}})$, considering here the columnwise median.
3. For each new estimate $(\hat{\mu}_k, \hat{\Sigma}_k)$:
   - Compute the Mahalanobis distance $\operatorname{MD}(x_i, \hat{\mu}_k, \hat{\Sigma}_k) = \sqrt{(x_i - \hat{\mu}_k)^T \, \hat{\Sigma}_k^{-1} \, (x_i - \hat{\mu}_k)}, \; i = 1, ..., n$ and choose the subset H_0 with the $\left\lceil \dfrac{n}{2} \right\rceil$ observations with smallest distance.
   - Compute $(\hat{\mu}_0, \hat{\Sigma}_0)$ based on H_0 and recalculate $\operatorname{MD}(x_i, \hat{\mu}_0, \hat{\Sigma}_0)$, choosing the h-subset with smallest distance and apply C-steps until convergence.
4. Retain the h-subset with smallest covariance determinant and obtain the final estimate $(\hat{\mu}, \hat{\Sigma})$.

The six initial estimates come from different methods, including estimators that are not discussed in this paper but can be seen in [2].

Proposition 2.5) The FAST-MCD algorithm is affine equivariant, but the DetMCD is not.

Despite the DetMCD not being fully affine equivariant, the deviation is small enough that it can be unconsidered. Hubert, Rosseeuw and Verdonck (see [11]), analysed thoroughly both algorithms and compared their performances. From it, we can list some properties that indicate when is it better to use each algorithm:

- For small to moderate number of dimensions (say $p \leqslant 10$), DetMCD is faster than FAST-MCD and equally robust. For higher dimensions (say $p > 10$), DetMCD is faster and even more robust.
- For minimally contaminated data sets, both algorithms have good efficiency. For highly contaminated data sets, DetMCD has a lower error for the location and scatter estimates.
- DetMCD does not depend on a random subset, making it a better option for computing the MCD estimates with different values of h.

2.3 Reweighted MCD

The MCD estimator has a rather low efficiency at the normal model. It increases as h increases but in exchange the breakdown value lowers. One way to increase the efficiency of the MCD estimator without compromising the breakdown value is to add a reweighting step to the estimator.

After a first iteration of the MCD, yielding the estimates $(\hat{\mu}_0, \hat{\Sigma}_0)$ of location and scatter, respectively, consider the sequence of scalars $\{w_i\}_{1 \leqslant i \leqslant n}$ where

$$w_i = \begin{cases} 1, \text{ if } \mathrm{MD}(x_i, \hat{\mu_0}, \hat{\Sigma}_0) \leqslant \sqrt{\chi^2_{p,0.975}} \\ 0, \text{ otherwise} \end{cases}$$

We can then calculate new estimates of location and scatter given by:

$$\hat{\mu}_{\text{new}} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \qquad \hat{\Sigma}_{\text{new}} = \frac{\sum_{i=1}^{n} w_i \left(x_i - \hat{\mu}_{\text{new}}\right)\left(x_i - \hat{\mu}_{\text{new}}\right)^T}{\left(\sum_{i=1}^{n} w_i\right) - 1}$$

Most algorithm implementations of the MCD estimator, including the one used for this paper, already execute a reweighing step before presenting the final estimates.

3 MVE estimator

The Minimum Volume Ellipsoid estimator was the first high-breakdown estimator of multivariate location and scatter to be used regularly in practice, preceding the MCD estimator.

Definition 3.1 (MVE estimator)

Given a p-variate n-data set, for a fixed h such that $\frac{n+p+1}{2} \leqslant h \leqslant n$, the MVE estimations of location and scatter are:

$$(\hat{\mu}, \hat{\Sigma}) = \underset{\mu,\Sigma}{\operatorname{argmin}} |\Sigma|$$

over all real μ and symmetric positive definite Σ that satisfy

$$\#\{i : d_i = \sqrt{(x_i - \mu)^T \, \Sigma^{-1} \, (x_i - \mu)} \leqslant \sqrt{\chi^2_{p,0.975}} \, \} \geqslant h$$

The MVE estimations are thus defined by the ellipsoid containing at least h observations for which the covariance matrix determinant is minimal

Intuitively, the main difference between the MCD and MVE estimators, is that the MVE estimator looks for an ellipsoid with minimal volume such that h points are either inside or in the boundary of the ellipsoid and has $p+1$ points in its boundary, while the MCD estimator only has the restriction of being minimal volume ellipsoid with h points inside or in its boundary.

3.1 Properties

Proposition 3.1) The MVE estimator is affine equivariant.

Proof: The proof of this proposition is analogous to the same proof for the MCD estimator. Given a data set X, consider the ellipsoid E containing h points of X with center $\hat{\mu}$ and scatter matrix $\hat{\Sigma}$ which corresponds to the MVE estimations of location and scatter.

Taking the function $f : \mathbb{R}^p \longrightarrow \mathbb{R}^p$ defined as $f(x) = Ax + b$, where A is a $p \times p$ non-singular matrix and b a p-dimensional vector, it is easy to see that f is continuous and therefore, the image $f(E)$ of the ellipsoid E will be an ellipsoid that contains the same h points as the original ellipsoid.

As we know from Proposition 2.1), $|\Sigma(AX + b)| = |A|^2 \, |\Sigma(X)|$ and so, if the pair $(\hat{\mu}, \hat{\Sigma})$ is the estimations location and scatter of the MVE estimator applied to the data set X, then $(A\hat{\mu} + b, A \, \hat{\Sigma} \, A^T)$ give the estimations for the data set $AX + b$, and we have affine equivariance.

Proposition 3.2) If the data is in general position, the MVE estimators of location and scatter will have a BV equal to

$$\varepsilon^*_n(T_\mu, X) = \varepsilon^*_n(T_\Sigma, X) = \frac{\min(n - h + 1, \, h - p)}{n}$$

As for the MCD estimator, using $\alpha = \dfrac{h}{n}$, we have that

$$\lim_{n\to\infty} \varepsilon^*_n(T_\mu, X) = \lim_{n\to\infty} \varepsilon^*_n(T_\Sigma, X) = \min(\alpha, \, 1 - \alpha)$$

and similarly, maximum BV is achieved when $h = \dfrac{n+p+1}{2}$

Proposition 3.3) The MVE estimator is not asymptotically normal (whereas the MCD estimator is) and has a low efficiency for finite samples.

This fact was studied by Davies in [10]. Like the MCD estimator, a reweighting step can be added to the estimator without compromising the breakdown value and increasing the efficiency for finite sampling.

4 Case analysis

In this chapter, there are three examples presented using the MCD and MVE estimators to better illustrate the usage of each estimator. All graphs and information about the data sets were worked in *RStudio* software.

4.1 The size of perches

A set of 56 perches were measured for their height (in cm) and weight (in g). The plot of the observations is shown below.

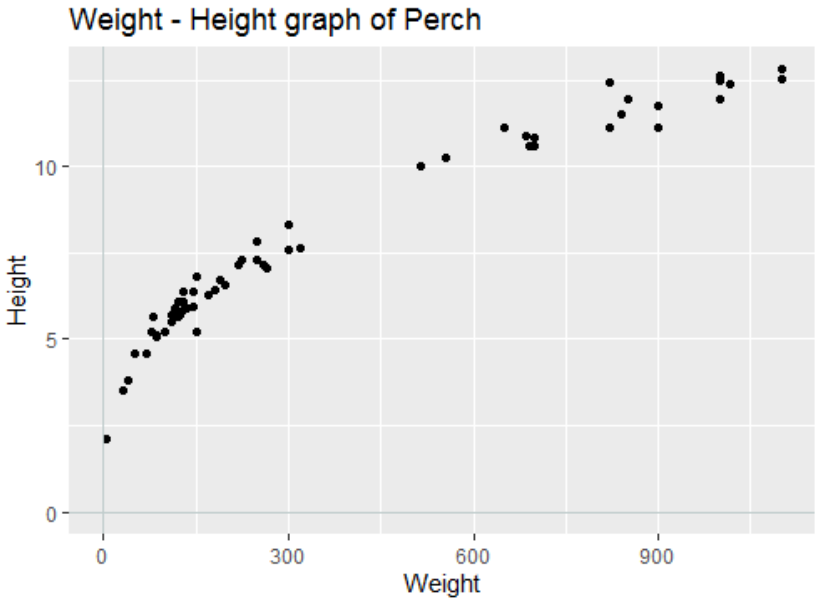


Figure 1: Set of observations of 56 perches

In order to estimate the parameters of location and scatter from this data set, there were used the classic mean and covariance matrix, the MCD and MVE estimators. The graph below shows for each method its tolerance ellipse and a coloured dot that represents the location parameter. In both the MCD and MVE estimators, it was used $\alpha = \frac{h}{n} = 0.5$

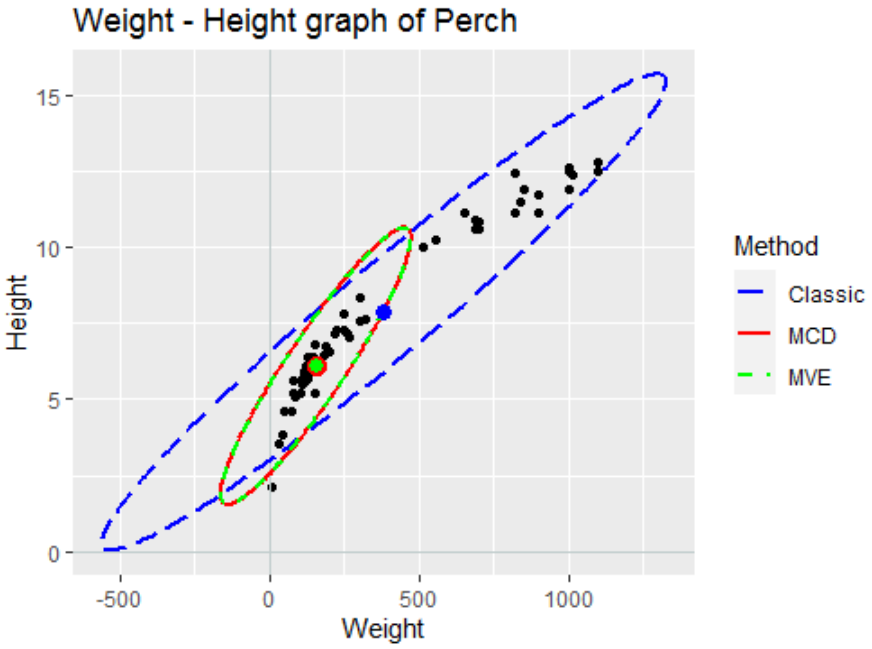


Figure 2: Tolerance ellipses for different methods

The figure above shows a perfect example of the discrepancy between the classical method and the MCD and MVE estimators. The MCD and MVE ellipses are overlapping, meaning they produced very similar estimates of location and scatter. As for the classical tolerance ellipse, the center is very different from the others and it actually only flags one observation as an outlier, where as MCD and MVE both flag 20 observations as outliers.

To show the effect of α when using the robust estimators, the graph below shows the tolerance ellipses and the center estimate when using the MCD estimator for $\alpha = 0.5$, $\alpha = 0,7$ and $\alpha = 0.9$.

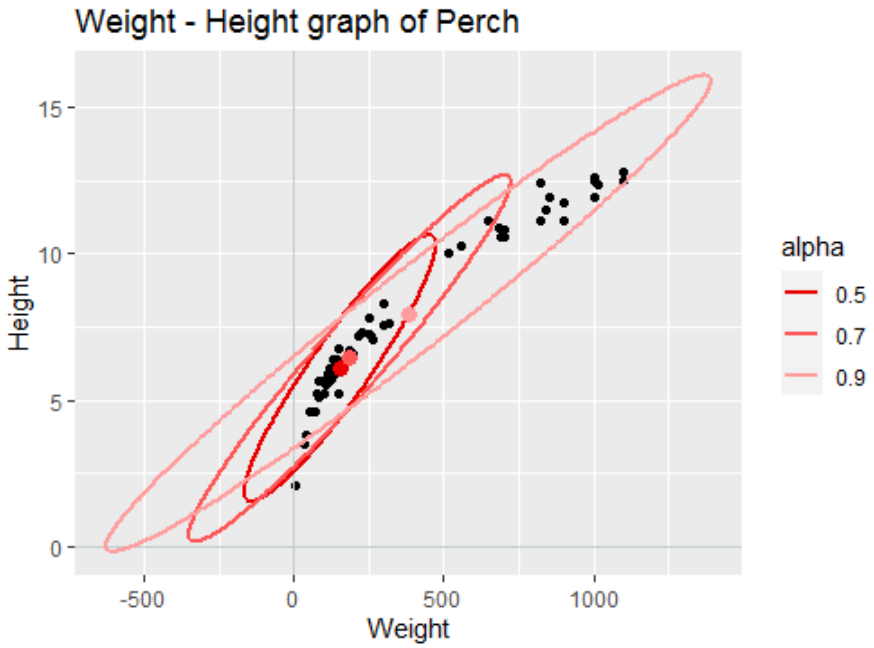


Figure 3: Tolerance ellipses for different values of α using MCD

Noticeably, the ellipse "grows" as the value of α increases in order to contain a larger portion of observations. Consequently, the location estimate (coloured dot) is shifting closer to the classical location estimate as α increases.

The tables below contain the information regarding the location estimate, the determinant of the scatter matrix and the number of flagged outliers.

	Method		
	Classic	**MCD**	**MVE**
$\hat{\mu}$	$\begin{pmatrix} 382.406 \\ 7.964 \end{pmatrix}$	$\begin{pmatrix} 154.431 \\ 6.091 \end{pmatrix}$	$\begin{pmatrix} 154.431 \\ 6.091 \end{pmatrix}$
$\log \lvert\hat{\Sigma}\rvert$	11.038	5.48	8.563
# outliers	1	20	20

	α		
	0.5	**0.7**	**0.9**
$\hat{\mu}$	$\begin{pmatrix} 154.431 \\ 6.091 \end{pmatrix}$	$\begin{pmatrix} 186.654 \\ 6.428 \end{pmatrix}$	$\begin{pmatrix} 382.406 \\ 7.964 \end{pmatrix}$
$\log \lvert\hat{\Sigma}\rvert$	5.48	8.121	10.15
# outliers	20	17	2

Table 1: Statistics about the estimates and outliers of each graph

Despite the fact that the MCD and MVE estimator have the same estimate of location and have almost identical tolerance ellipses, the determinant of the scatter matrix is significantly lower for the MCD estimator.

Even though the location estimate for the MCD when $\alpha = 0.9$ equals the classical method, the scatter matrix determinant is lower and it flags 1 more outlier. It makes sense that when α increases the MCD estimates get closer to the classical method because in the limit $\alpha = 1$, it would mean $h = n$ and then the MCD estimations of location and scatter would simply be the classic mean and sample covariance matrix (of all observations).

4.2 Protein expression in trisomic mice

In an experiment led by the University of Madrid and the University of Virginia (USA), it was studied the expression of 77 different proteins over 72 different mice. Each mouse was measured 15 times, totaling 1080 observations.

The mice were divided in 8 classes according to 3 different parameters: genotype (control mice or mice with Down Syndrome), behaviour (stimulated to learn, through context-shock treatment, or not, through shock-context) and what drug they were administrated during the experiment (saline or memantine).

Each class of mice is labeled as A-B-C, where A indicates the genotype, control (c) or trisomic (t), B indicates the behaviour, context-shock (CS) or shock-context (SC) and C indicates the administrated drug, saline (s) or memantine (m).

To analyze the effectiveness of the MCD and MVE estimators, there were considered two subsets, one with 4 proteins (NR1_N, NR2A_N, pNR2A_N, pNR2B_N) and the second with 40 proteins, which include the 4 proteins of the first subgroup. For each separate class, there were used the classic mean and covariance matrix, the MCD and MVE estimators (both using $\alpha = 0.5$) to determine the best estimates of location and scatter and detect outliers.

Figures 4 and 5 (pages 14 and 15) have the graphs that display the Mahalanobis distance for the different methods and groupings, for the 4 and 40 proteins subsets, respectively.

The table below has the number of detected outliers for each method. In each subset, the column "Separately" counts the total number of outliers among all classes.

	4 proteins		**40 proteins**	
	Separately	Conjoined	Separately	Conjoined
Classic	27	24	100	106
MCD	204	48	335	478
MVE	171	58	342	273

Table 2: Number of outliers for each method, grouping and protein subset

Observation 1) The 40 proteins subset registered a higher number of outliers compared to each corresponding graph from the 4 proteins subset.

Particularly when the classes were analyzed separately this is expected. Since each class has about 130-150 observations, the number of dimensions (e.g. proteins) is too high for how little observations there are.

Huber (see [12]) proved that the bias associated with the estimation of a sample covariance matrix of a p-variate n-data set is in the order of $O(p/n)$, and so for the 40 proteins subset and separate classes, we have $p = 40$ and $n \approx 150$, generating a big bias in the estimations of scatter. A recommended rule to avoid this is to work with data sets that have $n \geqslant 5p$.

Observation 2) The classic method registers almost the same number of outliers when working with separate and conjoined classes, for both protein subsets. The robust estimators registered more outliers when dealt with classes separately, except for the MCD estimator in the 40 proteins subset.

In the 4 protein subset, where n is significantly greater than p even in separate classes, we can disconsider the "error" effect of the number of dimensions. Since each class will have its own differences, it makes sense that the total number in separate classes is greater than the conjoined classes, because with conjoined classes the estimates are "normalized" over all of them instead of individually.

In the 40 protein subset, as seen before, the number of dimensions makes the propagation of small deviations in each dimension be very impactful when calculating the Mahalanobis distance. Because $\alpha = 0.5$ is the lowest possible value for these robust estimators, when working with conjoined classes it means over 500 observations are not being accounted for the estimates, which can explain why the MCD estimator registered a higher number of outliers in conjoined classes compared to separate.

The table below has the estimates of location (for **conjoined** classes) of each protein of the first subset and the corresponding value of those four proteins in the location estimate of the second subset. It also shows the determinant of the scatter matrix estimate of each case.

	4 proteins			**40 proteins**		
	Classic	**MCD**	**MVE**	**Classic**	**MCD**	**MVE**
NR1_N	2.2744	2.2972	2.2704	2.2950	2.2540	2.2682
NR2A_N	3.7375	3.8439	3.7331	3.8323	3.7514	3.7722
pNR2A_N	0.6976	0.7269	0.6998	0.7293	0.7361	0.7212
pNR2B_N	1.5369	1.5619	1.5389	1.5620	1.5463	1.5482
$\log \lvert\hat{\Sigma}\rvert$	-12.15	-12.83	-12.84	-239.09	-259.58	-253.81

Table 3: Location estimates and scatter matrix determinant for each method and subset in conjoined classes

Observation 3) In the 4 proteins subset, the MVE location estimate is closer to the classic method than to the MCD. In the 40 proteins subset, the MCD and MVE location estimates are much closer to each other and different from the classic method.

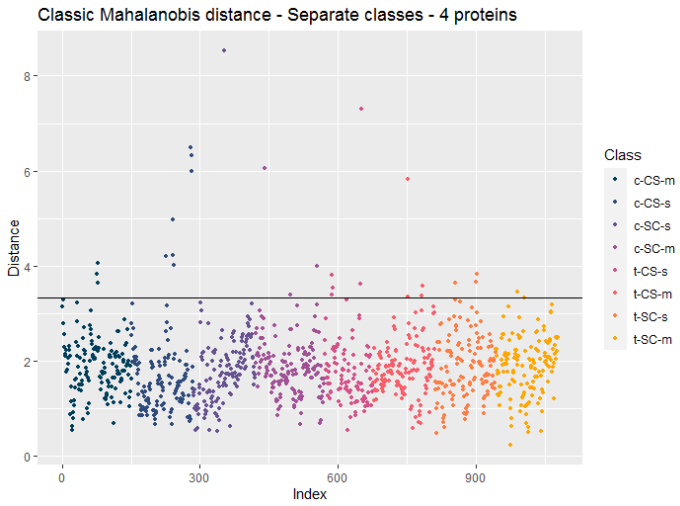
This provides an example to the difference between both estimators. Despite having similar premises of finding a certain minimal volume ellipsoid, they can in fact produce different estimates.

Another important fact to note is that, as mentioned in the MCD estimator section, the exact algorithm for the MCD estimator is very computationally expensive for large data sets. Therefore, most available softwares use FAST-MCD or DetMCD for these calculations (in this particular example, it was used DetMCD). Because it doesn't run through all h-subsets, it is not guaranteed to reach the global minimum.

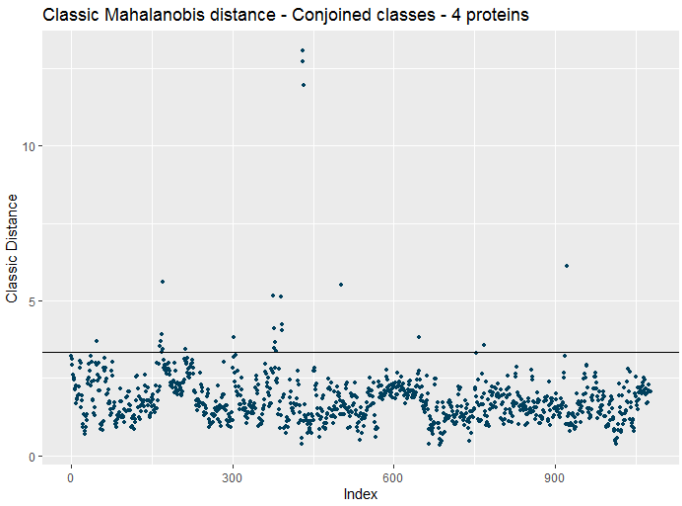
Observation 4) In the 4 proteins subset, the scatter matrix determinant reduced in about half from the classic method to the MCD and MVE methods. In the 40 protein subset, the scatter matrix determinant reduced in about 10^9 times from the classic to the MCD, and about 10^7 from the classic to the MVE method.

The difference between this rate of reduction can be explained by the effect of the number of dimensions. Thinking of the determinant as being proportional to the volume of the ellipsoid defined by it and knowing that the determinant of a matrix can be calculated as the product of all its eigenvalues, in a higher dimensional setting it is expected that the ellipsoid gets more "contracted". Given that the measurements of the data set are fairly small and looking at the values of the determinant for the classic method, we can assume part of the eigenvalues get smaller than 1, which in a product with more terms generates an even smaller result.

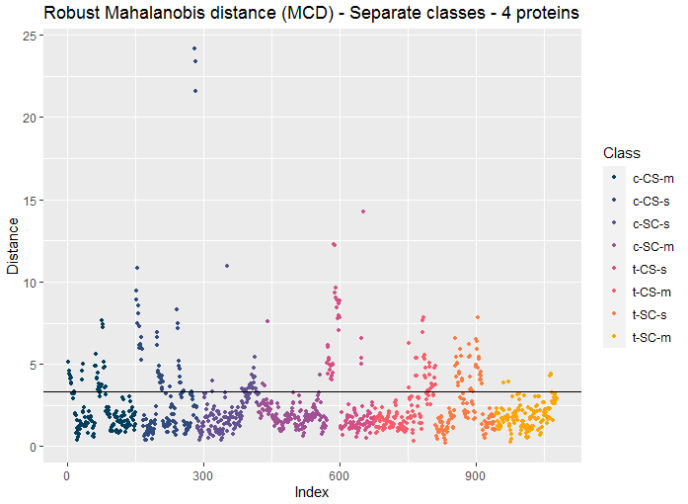
The images below show the Mahalanobis distance plot for each method (Classic, MCD and MVE) and analyzing classes separately (each class with its own estimate of location and scatter) and conjoined (one common estimate) for the **4 proteins subset**.



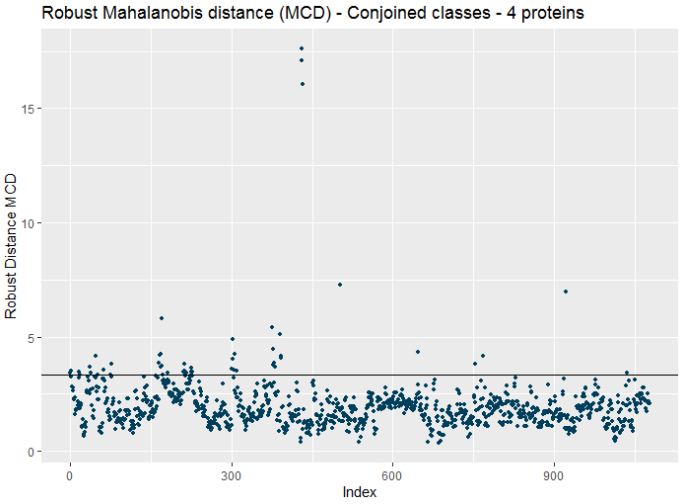
(a) Classic - Separate classes



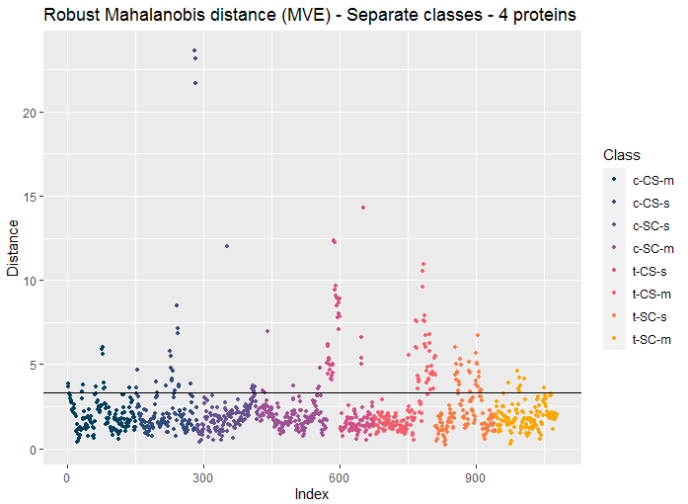
(b) Classic - Conjoined classes



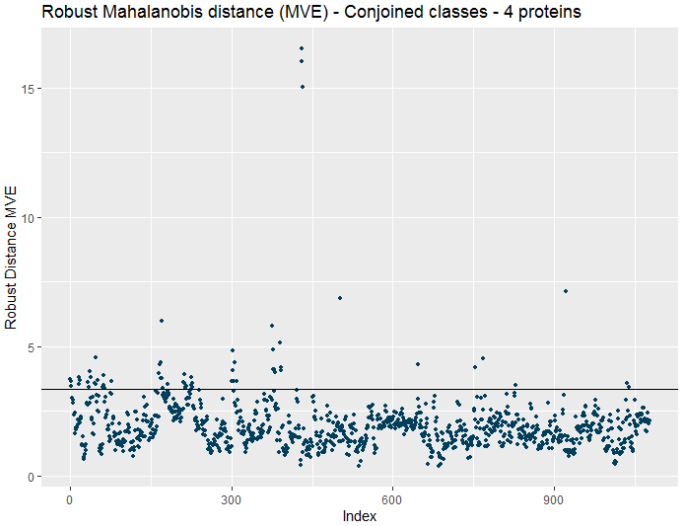
(c) Robust MCD - Separate classes



(d) Robust MCD - Conjoined classes



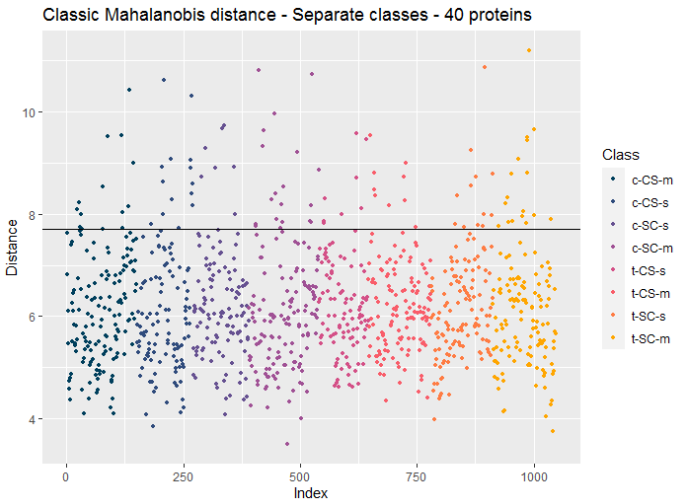
(e) Robust MVE - Separate classes



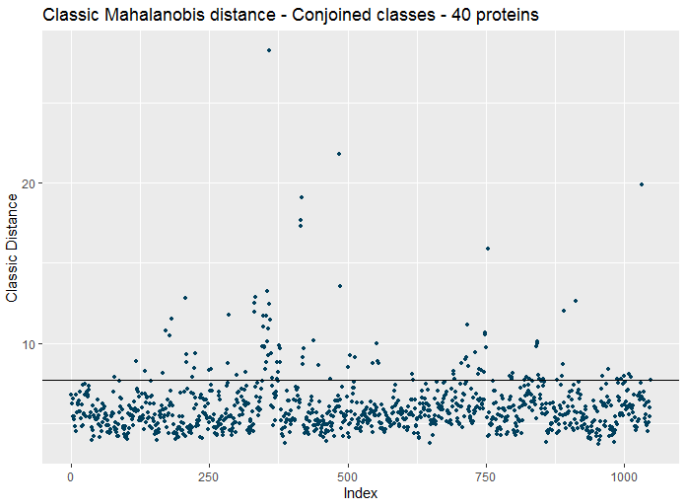
(f) Robust MVE - Conjoined classes

Figure 4: Mahalanobis distance for different methods in separate and conjoined classes (4 proteins)

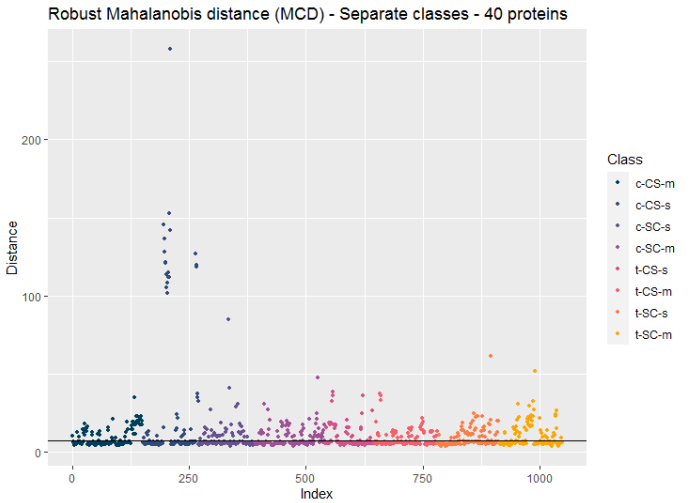
The images below show the Mahalanobis distance plot for each method (Classic, MCD and MVE) and analyzing classes separately (each class with its own estimate of location and scatter) and conjoined (one common estimate) for the **40 proteins subset**.



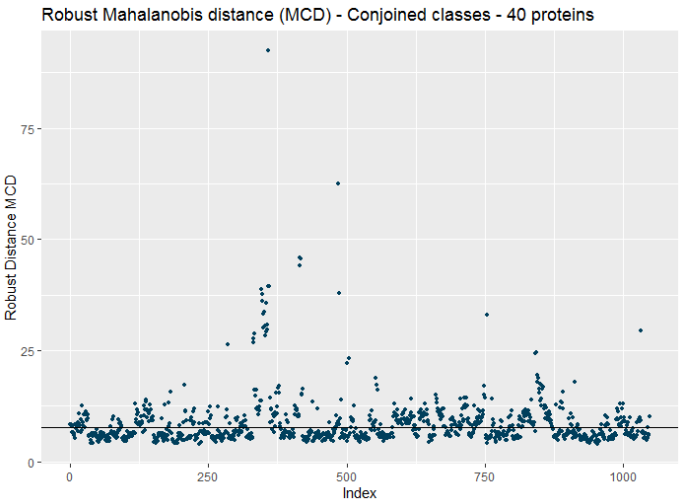
(a) Classic - Separate classes



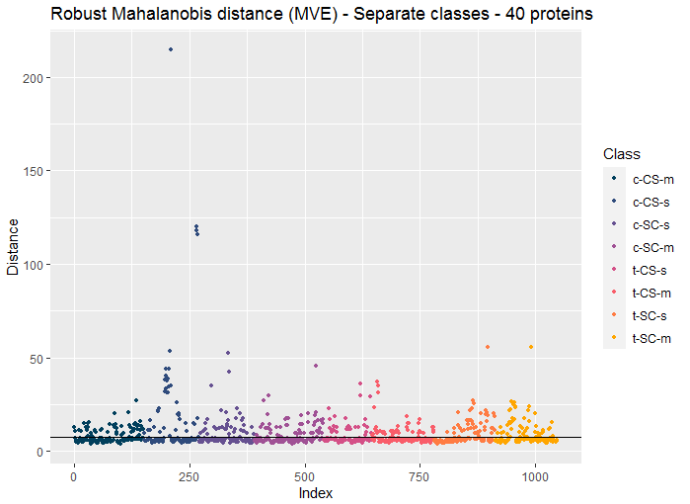
(b) Classic - Conjoined classes



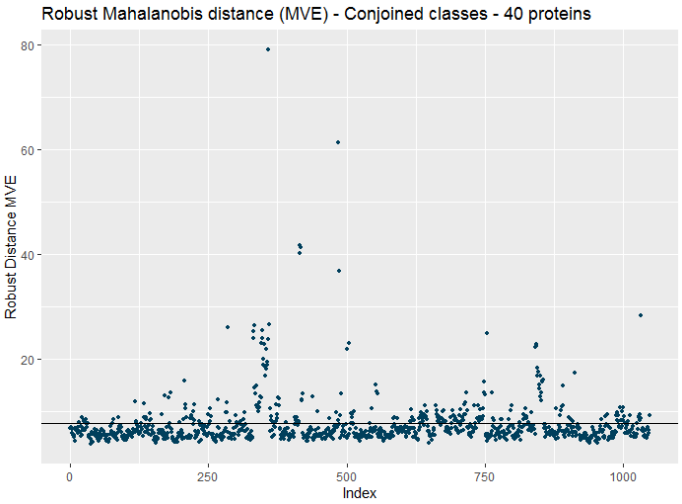
(c) Robust MCD - Separate classes



(d) Robust MCD - Conjoined classes



(e) Robust MVE - Separate classes



(f) Robust MVE - Conjoined classes

Figure 5: Mahalanobis distance for different methods in separate and conjoined classes (40 proteins)

4.3 The Hit Potential Equation

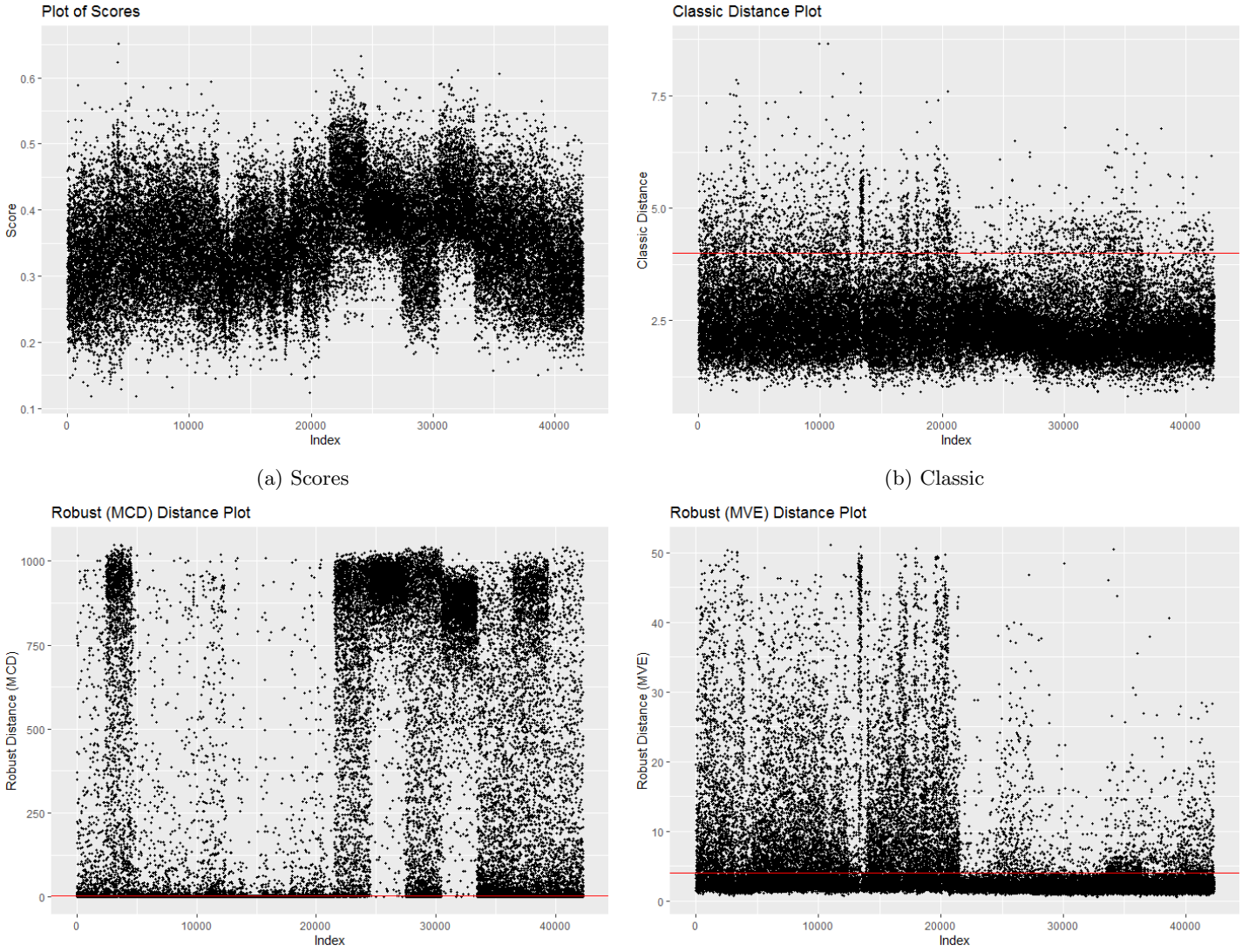
In a crossover between Machine Learning and music, the University of Bristol developed an algorithm that can, in theory, predict with 60% accuracy whether or not a new song can become widely popular (so called "hit"). As a result, they developed The Hit Potential Equation, which gives a song a score based on different attributes. The equation has this format:

$$\text{Score} = w_1 f_1 + w_2 f_2 + w_3 f_3 + ...$$

The w_i are weights given to i^{th} attribute of the song in analysis and the f_i represent a scalar assigned to the same attribute which compares the song with past hits in order to see if it follows the same trend as current popular songs. The higher the score, the more likely it is for the song to become a hit.

In this section, we will be analysing a data set with a total of 42305 songs over 7 different attributes: danceability, energy, speechiness, acousticness, instrumentalness, liveness and valence. Each attribute's weight w_i is rated as a real number from 0 to 1, and for simplicity, we will assume $f_i = 1/7, \forall i$, making our reworked Hit Potential Equation be the mean of the scores of the attributes.

The figure below shows the scores of each song of data set, as well as the Mahalanobis distance for the classic, MCD and MVE (both with $\alpha = 0.5$) estimates.



(a) Scores

(b) Classic

(c) MCD

(d) MVE

Figure 6: Scores and Mahalanobis distances plot

The table below has the registered number of outliers as well as the location estimate of each attribute for each method.

	# outliers
Classic	2588
MCD	21809
MVE	12957

	Danc.	Energy	Speec.	Acoust.	Instrum.	Liveness	Valence
Classic	0.639	0.762	0.136	0.096	0.283	0.214	0.357
MCD	0.663	0.696	0.184	0.146	0.00026	0.209	0.426
MVE	0.626	0.822	0.103	0.0128	0.372	0.223	0.332

Table 4: Number of outliers detected and location estimates for each method

Analyzing the graphs and the table, there are 3 noticeable things to point out regarding the MCD estimations:

- The maximum value of the Mahalanobis distance is over 1000 for the MCD, whereas it is about 8 for the classic method and a slightly over 50 for the MVE.
- The Instrumentalness estimate of location is surprisingly low compared to the other methods.
- The MCD flagged over 50% of the observations as outliers.

Immediately we know these estimates for the MCD are not accurate. As seen in Proposition 2.3), the BV of the MCD estimator cannot exceed 50%, so it could never flag more than half of the observations as outliers. To explain this odd behaviour, the graph below represents the histogram of the Instrumentalness attribute.

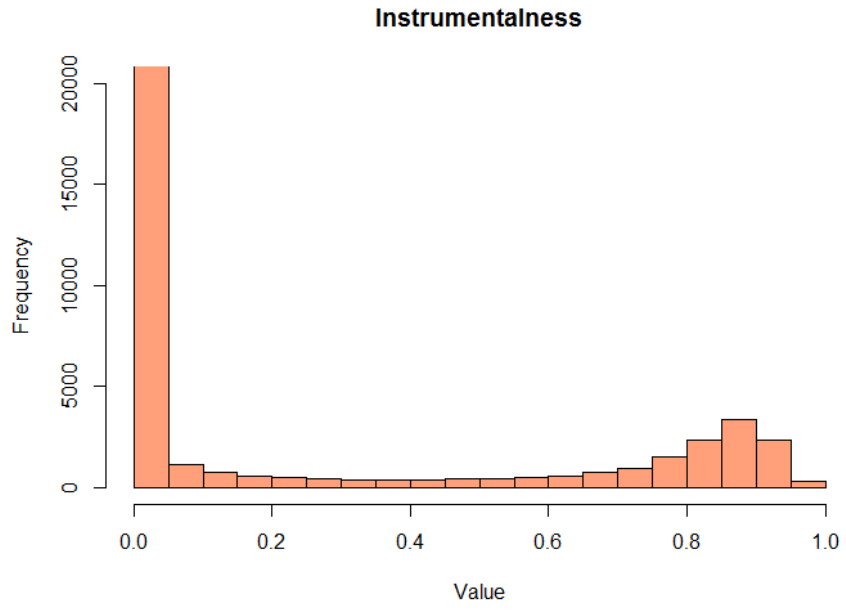


Figure 7: Histogram of Instrumentalness

Nearly 24000 (over half the observations) songs of the dataset have an Instrumentalness rated below 0.05. Because the MCD was used with $\alpha = 0.5$, it means that the location estimate of the MCD was "dragged down" to a cluster of songs with very low Instrumentalness.

More importantly, this histogram shows that this data (or at least this variable) cannot be fit into a normal model, as we've previously been doing. As such, the MCD (and even the MVE) are not useful under these circumstances. Even when using the robust estimators with a higher value of α, it still results in very distorted estimates.

5 Conclusion

In this paper, we started with an introduction to robust statistics in a multivariate setting. There were presented the notions of outliers, estimators and some robustness measures. The rest of the paper was dedicated to studying two of the most widely known and used robust estimators: the Minimum Covariance Determinant (MCD) estimator and Minimum Volume Ellipsoid (MVE). Firstly in a theoretical approach, analyzing some of its properties and what makes them such robust estimators, and afterwards we were able to see them in practical cases, where we verified all the theoretical statements regarding the estimators.

It was made a proper comparison between both robust estimators and classic method of determination of location and scatter. It was also seen the limitations of the estimators, mainly the computational difficulties of performing the exact algorithm, hence the introduction of new and more feasible algorithms, and the cautions to have regarding the data sets.

The detection of outliers in Data Analysis serves as the first step to many scientific areas as Machine Learning, Linear Regression or other areas of Engineering and Economics, and estimators like the MCD and MVE are very useful tools to accurately read large multivariate data as they have proven to effectively detect and remove outliers.

References

[1] P. Rousseeuw, Robust Statistics Part 1: Introduction and univariate data, LARS-IASC School, May 2019

[2] P. Rousseeuw, Robust Statistics Part 2: Multivariate location and scatter, LARS-IASC School, May 2019

[3] P. Rousseeuw and M. Hubert, Anomaly detection by robust statistics, WIREs Data Mining Knowl Discov 2018

[4] M. Hubert and M. Debruyne, Minimum covariance determinant, WIREs Comp Stat 2010 2 36–43

[5] C. Croux and G. Haesbroeck. Influence function and efficiency of the Minimum Covariance Determinant scatter matrix estimator. Journal of Multivariate Analysis, 71: 161–190, 1999.

[6] S. Van Aelst and P. Rousseeuw, Minimum volume ellipsoid, Wiley Interdisciplinary Reviews Computational Statistics, July 2009

[7] P. Rousseeuw and K. Van Driessen, A Fast Algorithm for the Minimum Covariance Determinant Estimator, TECHNOMETRICS, Vol. 41, No. 3, August 1999

[8] M. Thill, The Relationship between the Mahalanobis Distance and the Chi-Squared Distribution, September 2017

[9] R. Grübel, A Minimal Characterization of the Covariance Matrix, Metrika, Volume 35, 1988

[10] Davies L., The asymptotics of Rousseeuw's Mini-mum Volume Ellipsoid estimator. Ann Stat 1992, 20:1828–1843

[11] M. Hubert, P. Rosseeuw and T. Verdonck, A Deterministic Algorithm for Robust Location and Scatter, Journal of Computational and Graphical Statistics, Volume 21, Number 3, Pages 618–637, 2012

[12] P. Huber, Robust Regression: Asymptotics, Conjectures and Monte Carlo, The Annals of Statistics Vol.1 No.5, 1973