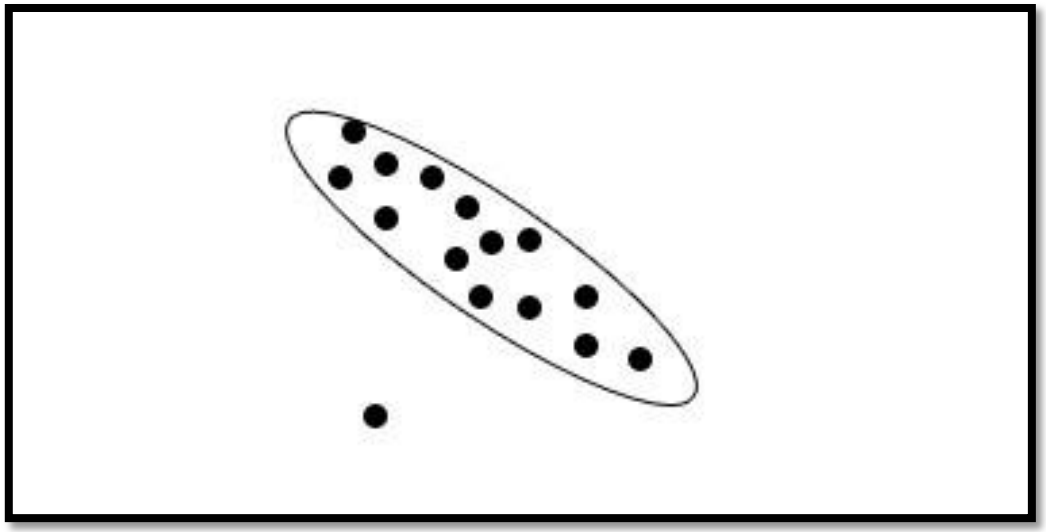
Multivariate outliers detection

Lavrentyeva Galina

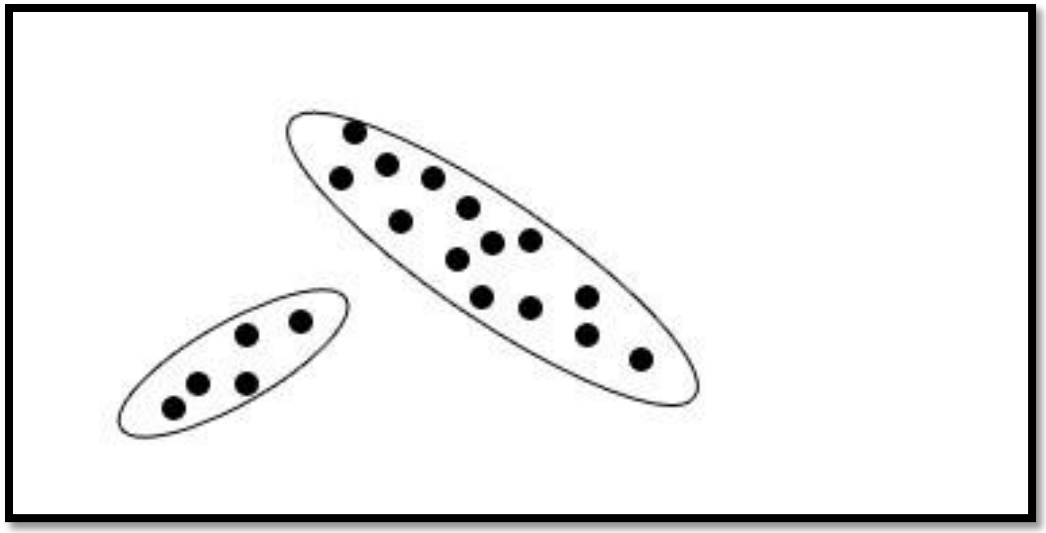
October 30, 2012

Types of outliers

- Isolated

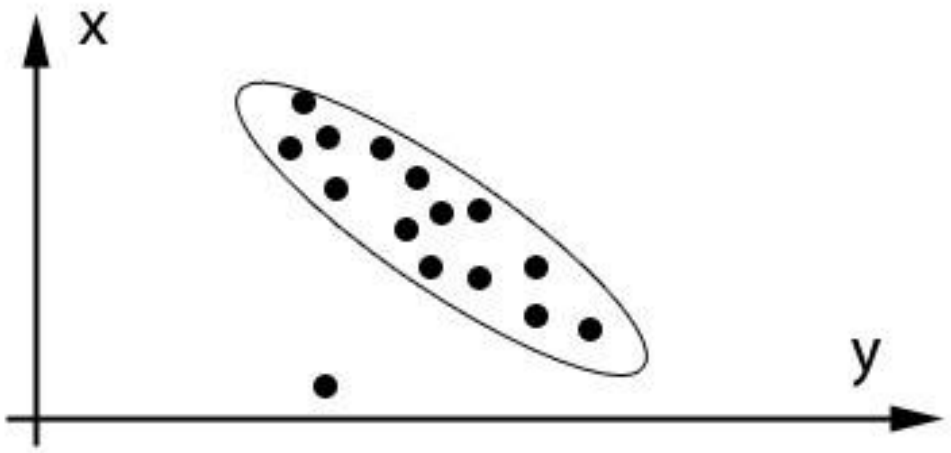


- Clusters



Multivariate analysis

Only multivariate analysis is performed, and the interactions among different variables are considered.



Masking effect

One outlier masks a second outlier, if the second outlier can be considered as an outlier only by itself, but not in the presence of the first outlier.

Swamping effect

One outlier swamps a second observation, if it can be considered as an outlier only under the presence of the first one.

Statistical methods. Mahalanobis distances

Given: n observations from p-dimentinal dataset with sample mean vector $\overline{x_n}$ and covarince matrix V_n.

Outlier detection

For each point i = 1, 2…n:

$$M_i = \sqrt{\sum_{1}^{n} (x_i - \overline{x_n})' {V_n}^{-1} (x_i - \overline{x_n})}$$

Observations with large Mahalanobis distances are indicated as outliers.

(!) Work well only when outliers are few and isolated (masking, swampinf effects).

Robust estimates

Emperical mean -> medians

Covariance ->

- MCD (Minimum Covariance Determinant)
- S-estimators
- MVE (Minimum Volume Ellipsoid)

Data-mining methods

Often non-parametric, thus, do not assume an underlying generating model for the data.

- distance-based methods
- clustering methods
- spatial methods

Distance-based

Def.(Knorr and Ng, 1997): An observation is defined as a distance - based outlier if at least a fraction β of the observations in the dataset are further than r from it.

Acuna and Rodriguez (2004): such definition raises difficulties, such as the determination of r and the lack of a ranking for the outliers.

The time complexity of the algorithm: $O(pn^2)$, where p is the number of features and n is the sample size.

Not adequate def. if:

- very large dataset
- dataset has both dense and sparse regions

Distance-based (2)

Def. (Ramaswamy et al., 2000):

given two integers v and i (v, i), outliers are defined to be the top i sorted observations having the largest distance to their v-th nearest neighbor.

Outliers are observations that have a large average distance to the v-th nearest neighbors.

Longer time to be calculated.

Clustering

Cluster-based methods consider a cluster of small sizes, including the size of one observation, as clustered outliers.

In most cases, the outlier detection criteria are implicit and cannot easily be inferred from the clustering procedures.

Further inspection of detected outliers is necessary for good data analysis.

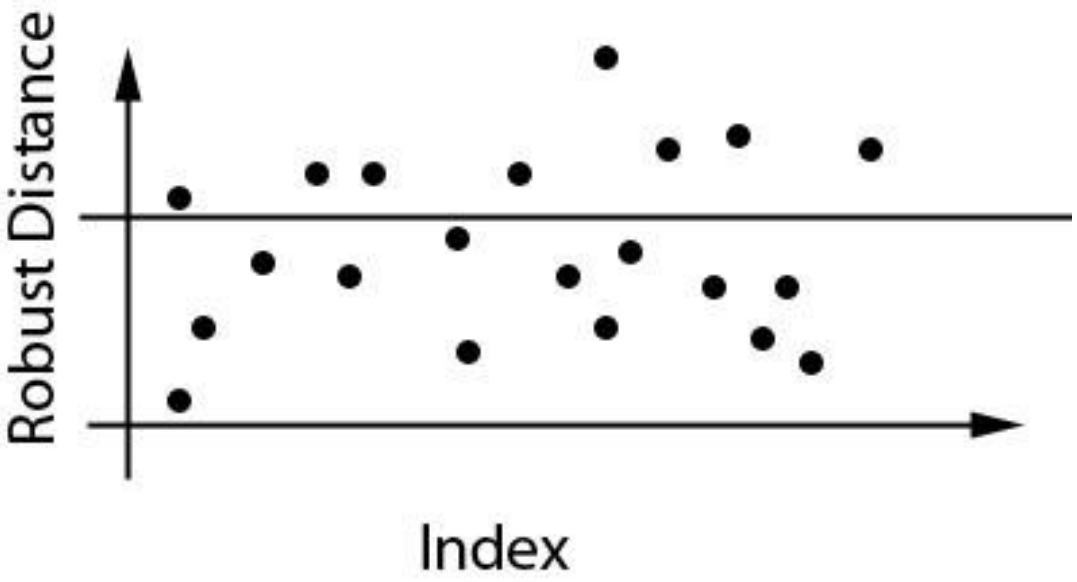
Clustering (2)

In this situation structure would imply that the outliers form a cluster of their own, clearly separated from the other cluster (majority of the data).

Underlying distribution = mixture distribution

Clustering (3)

Robust distances do not reveal the structure of the outliers.



Robust distances based on MCD for simulated data (clustered outliers)

Clustering (4)

Def.: Group of outliers can be regarded as a separate cluster whenever they can approximately be covered by an ellipsoid which does not intersect the ellipsoid covering the data majority points.

$Def.$ ~existence of a separating hyperplane between data majority and outliers.

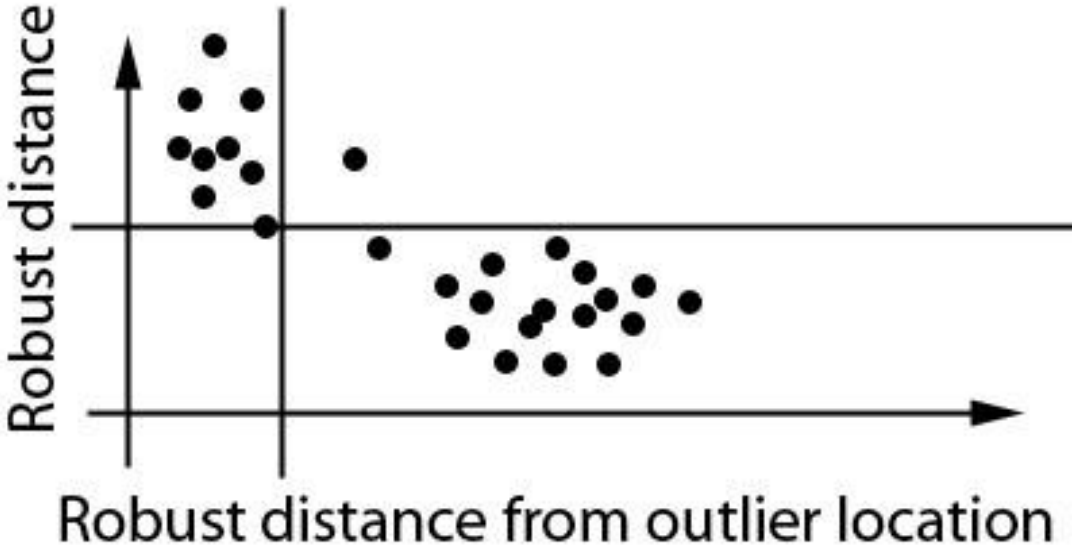
Distance versus distance

$\widehat{\mu_1}, \widehat{V_1}$ - robust estimates for the location and covariance for dataset X_n

1. Defining the data majority by Mahalanobis distances with $\widehat{\mu_1}, \widehat{V_1}$.
2. Points that fall outside are labeled as outliers.
3. Computation of $\widehat{\mu_2}, \widehat{V_2}$ for the outliers subset.
4. Defining the outlier majority by by Mahalanobis distances with $\widehat{\mu_2}, \widehat{V_2}$

Distance versus distance for simulated data

Plotting distances $d_i\left(\widehat{\mu_1}, \widehat{V_1}\right)$ versus $d_i\left(\widehat{\mu_2}, \widehat{V_2}\right)$ is more informative in this case.



Spatial

Spatial methods define a spatial outlier as a spatially referenced object whose non-spatial attribute values are significantly different from the values of its neighborhood.

Spatial methods:

- Quantitative methods (provide tests to distinguish spatial outliers from the remainder of data)
- Graphical methods (are based on visualization of spatial data which highlights spatial outliers)

Separation index

Can be computed for two clusters of observations and measures the scale of the sparse area between them. It is measured in all possible univariate projections.

$$SI = \max_{a} J(a)$$

$$J(a) = \frac{a^t(\widehat{\mu_2} - \widehat{\mu_1}) - z_{\frac{\alpha}{2}}(\sqrt{a^t\widehat{V_1}a} + \sqrt{a^t\widehat{V_2}a})}{a^t(\widehat{\mu_2} - \widehat{\mu_1}) + z_{\frac{\alpha}{2}}(\sqrt{a^t\widehat{V_1}a} + \sqrt{a^t\widehat{V_2}a})}$$

$$SI \in [-1; 1]$$

Future plans

- Bivariate linear projections
- Using robust FQ_n-estimators

$$FQ_n = MAD_n\left(1 - \frac{Z_0 - n/\sqrt{2}}{Z_2}\right)$$

$$Z_k = \sum_{i=1}^{n} {u_i}^k \, e^{-{u_i}^2/2} \qquad u_i = \frac{x_i - medx_i}{MAD_n}$$

$$MAD_n = 1.4826 \cdot med|x - medx|$$

Future plans (2)

- Check the accuracy of the proposed methods for real data and simulated data with noise

References

1. Barnett V., Lewis T. (1994). Outliers in Statistical Data. Wiley, New York.
2. Becker C., Gather U. (1999). “the masking breakdown point of multivariate identification rules”, Journal of the American Statistical association, 94, 947-955.
3. Willems G., Joe H., Zamar R. “Diagnosing multivariate outliers dectected by robust estimators”

Thank you!