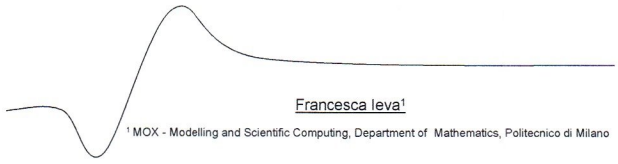
1. NONPARAMETRIC MULTIVARIATE EXPLORATION

Multivariate case:
what can we do if we know nothing about the distribution of the data? Depth.
We can use depth for:

- ranking objects in multivariate space
- graphical inspection of multivariate distribution
- performing estimations (location, dispersion, shape)

Depth Measures

Nonparametric Statistics
AA 2020-2021



Francesca Ieva[1]

[1] MOX - Modelling and Scientific Computing, Department of Mathematics, Politecnico di Milano

Outline

Setting

- This work aims at exploring the concept of depth in Statistics and at showing its usefulness in practice.
- Initially the concept of data depth is introduced. This concept is very important because it leads to a natural center-outward ordering of sample points in multivariate data sets.
- This notion of order in multivariate data sets enlarges the field of applications of multivariate analysis, since it allows the extension of univariate concepts based on order to the field of multivariate analysis, in particular it opens the possibility of nonparametric methods to be used in multivariate data analysis.
- Other challenges within the topic of data depth are the computational implementation of the depth measures and the graphical representation of the data depth.

very useful when we want to order objects (all depth theory is based on rankings)

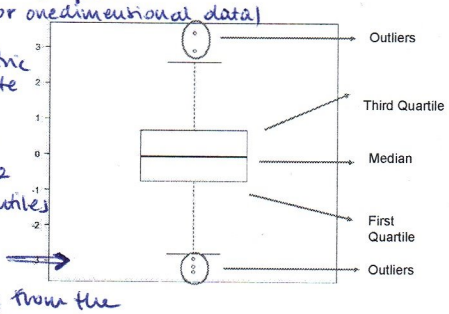
Setting

Boxplot:
(for onedimensional data)

main nonparametric tool for univariate data

this is already based on "depth": we put our data in line, we find the quantiles and then we find the median, Q_1, Q_3, ...

⇒ we have an ordering from the more central point to the most outlying one



Rank data according to their depth

Depths measure

Outlyingness

Centrality

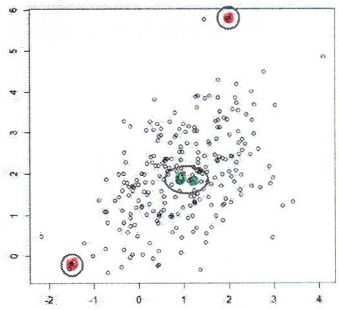
Outlyingness

What if the dimension of the data grows? p=2 ;

Green data are more central than the red ones, but which green is more central? which red is more outlying? It's not as intuitive as in the univariate case.
(we loose the "complete order" of the space)

Setting

Data generated according to a multivariate distribution with law P_X



- Green data are deeper than **red** ones
- More difficult to rank the two green points or the two red ones
- No total order in R^d when d>1

Aim: rank data

⇩

Need: depth measure in R^d

Why Nonparametrics

for example we want to describe data according to the first moment (mean), or we want to compute the variance (second moment),..
However, in multivariate settings the description based on moments is not optimal.

- When dealing with multivariate distributions, the commonly adopted method for obtaining multivariate distributional characteristics has been a straightforward extension of the moment approach in the univariate case. More specifically, the location, scale, skewness and kurtosis are defined, respectively, in terms of moments => matrix or vector forms of outputs which are hard to grasp conceptually and graphically.
  Note: this approach is not even be applicable if the moments do not exist.
- In Liu (1999) approach, these characteristics and their corresponding descriptive statistics are defined as functionals of data depth. They can than be displayed as simple graphs on the plane and be easily visualized.
- In general, the depth-based methodology may be viewed in part as a **multivariate generalization of standard univariate rank methods**.
- Difference in the two ways of ranking: the ranking in the **univariate case** is a *ranking* from the smallest to the largest, while our **multivariate one** is a *center-outward ranking* induced by depth.

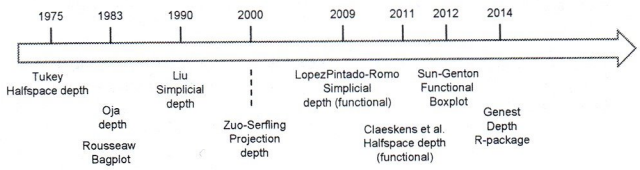
New approach: let's describe the data not with moments but with depts.
(= do not assume any distributional assumption below)

RANK and DEPTHS have no assumptions on the distributions of the data.

- 1D rank: ordering from the minimum to the max.
- 2D/nD rank: "ordering" from the more central to the more outlying

A bit of history

- The notion of data depth was proposed by Tukey (1975) as a graphical tool for visualizing bivariate data sets, and has since been extended to the multivariate case (Donoho and Gasko, 1992).
- The depth of a point relative to a given data set measures how deep that point lies in the data cloud. The data depth concept provides center-outward ordering of points in any dimension and leads to a new non-parametric multivariate statistical analysis in which **no distributional assumptions** are needed.



Application fields

- Nonparametric description of multivariate distributions (Liu et al.,1999; Serfling, 2004; and Wang and Serfling, 2005)
- Outlier identification in multivariate setting (Serfling, 2006; and Zhang, 2002)
- Depth-based classification and clustering (Ruts and Rousseeuw, 1996; Christmann, 2002; and Jörnsten, 2004)
- Rank and sign tests (Brown and Hettmansperger, 1989; Hettmansperger et al., 1992; and Hettmansperger and Oja, 1994)
- Multivariate density estimation (Fraiman et al., 1997)
- Data based linear regression (Rousseeuw and Hubert, 1999)
- Outlier identification in functional setting (Sun & Genton, 2011)
- Dimensional reduction for functional data (Tarabelloni et al. 2015, Ieva and Paganoni 2020)

The computational issue

depth theory is strongly based on pairing and geometry (and so it's connected with visualization), so, as dimension grows a lot of computational problems arise

- For the concept of data depth to be really useful, its computation has to be **possible and efficient**.
- A collaborative effort of statisticians and computational geometers is the foundation for ongoing research on this subject.
- Although there are already some satisfactory algorithms for computing depth functions in the bivariate case, there is still much work to do in higher dimensions (rarely dimensions higher than 2 or 3 is considered).

- ***depth*** - Genest, M., Masse, J.C., Plante, J.F. (2012). depth: Depth functions tools for multivariate analysis, http://CRAN.R-project.org/package=depth
- **DepthProc** - Kosiorowski D, Zawadzki Z (2020). *DepthProc An R Package for Robust Exploration of Multidimensional Economic Phenomena.*
- ***aplpack*** – Wolf, H.P., Bielefeld, U. (2014) aplpack: Another Plot PACKage: stem.leaf, bagplot, faces, spin3R, plotsummary, plothulls, and some slider functions, http://CRAN.R-project.org/package=aplpack) sono infatti 2012 - 2014 e poi
- ***roahd*** – Tarabelloni et al. (2019). roahd: RObust Analysis of High Dimensional Data http://CRAN.R-project.org/package=roahd

only 4 packages useful and comprehensive

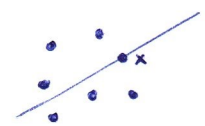
Depth Measures

- Let F be a cumulative distribution function associated to a probability distribution in R^d, $d \geq 1$. Unless stated otherwise, we assume that F is absolutely continuous and that $\{x_1, \dots, x_n\}$ is a random sample from F.
- Each sample point x_i is viewed as a $dx1$ column vector.
- A data depth is a way of measuring how deep (or central) a given point $\boldsymbol{x} \in R^d$ is w.r.t. F or w.r.t. a given data cloud $\{x_1, \dots, x_n\}$.
- Many definitions of depths are possible:
  - Half Space or Tukey depth
  - Simplicial or Liu depth
  - Oja depth
  - Mahalanobis depth
  - Convex Hull Peeling depth

  } most common
- **Note**: in nonparametric setting, we deal with infinite dimensional spaces, whereas in parametric setting we focus on a finite (tipically low) dimensional subspace, i.e., the space identified by the parameters

1. Half-space (or Tukey) Depth [HD]

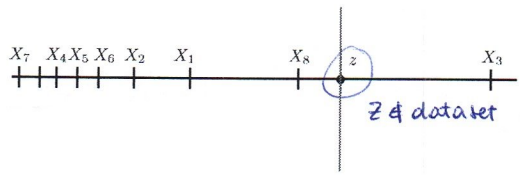
- Location depth of a point point $\boldsymbol{x} \in R^d$ relative to a d-dimensional data set is defined as the smallest number of data points in a closed halfspace with boundary passing through **x**.
- In the univariate case is easy to see that the depth of a point **x** is given by $\min\{ \#\{x_i \leq x\} ; \#\{x_i \geq x\}\}$ and the median is the point (or points) with maximal depth.
- Note that it might not be unique.

⇒ 

what is the depth of **x**? we consider a subspace of dimension d-1 (in this case 1) passing through x. In so doing we divide the data cloud in 2 parts: we count how many data we have in each part and we define the minimum of the two as the depth of **x**. In this case: depth(x) = 3

Consider the univariate case:

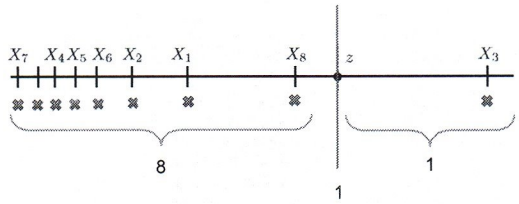
Half-space (or Tukey) Depth [HD]



Notice that the order is not $X_1, \dots, X_8$. For instance, X_2 is the second sample but we ordered the samples by their values.

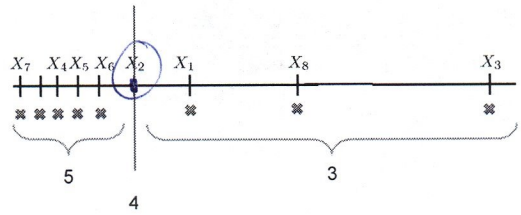
Half-space (or Tukey) Depth [HD]

We count how many points are there on the right and how many on the left:



This operation can be done on every point observed: (↓)

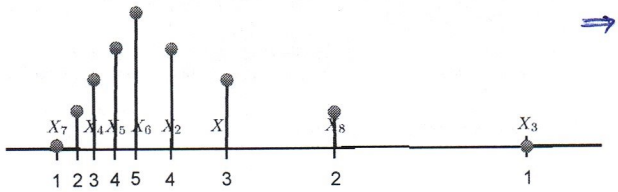
Half-space (or Tukey) Depth [HD]



5, 3 are calculated without X_2: if we add X_2 to the left we obtain 6, on the right 4
⟹ we take the minimum
⟹ depth(X_2) = 4

Half-space (or Tukey) Depth [HD]

Doing it for all X_i:



⇒ we have now points with higher depth and lower. We want something like that in higher dimensional cases

Half-space (or Tukey) Depth [HD]

- Location depth of a point point $\boldsymbol{x} \in R^d$ relative to a d-dimensional data set is defined as the smallest number of data points in a closed halfspace with boundary passing through **x**.
- In the univariate case is easy to see that the depth of a point **x** is given by $\min\{ \#\{x_i \leq x\} ; \#\{x_i \geq x\}\}$ and the median is the point (or points) with maximal depth.
- Note that it might not be unique.
- In the multivariate case, the notion of median can be generalized, being the point with maximal depth. This multivariate median is called the Tukey median

Half-space (or Tukey) Depth [HD]

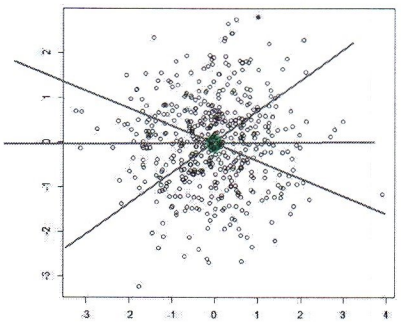
- The *half-space depth* [Hodges (1955), Tukey (1975)] at x w.r.t. F is defined to be

$$HD(F; x) = \inf_{H} \{P(H): H \text{ is a closed half-space in } \mathbb{R}^d \text{ and } x \in H\}$$

P(H)

- The sample version of $HD(F; x)$ is $HD(F_n; x)$. F_n denotes the empirical distribution of the sample $\{x_1, \ldots, x_n\}$.
- The half-space depth is sometimes also referred to as the Tukey depth in the literature.

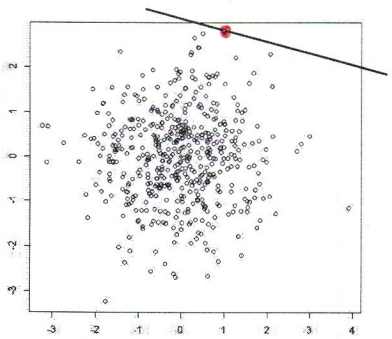
Half-space (or Tukey) Depth [HD]



We procede as we previously presented. Since we have to keep the infimum, we have to analyze all the possible subspaces passing through x ⇒ all the possible lines passing throug x.

How costly is it?
we have to do it for every point!

Half-space (or Tukey) Depth [HD]



For some points it's easier : for the red one we can (fastly) find a line such that on one part of the space there are no points ⇒ obviously that's the infimum!
(0 is the infimum)

However in the general case, this procedure is very costly.
But once this long procedure is over we can rank all the point (from those having higher depth to those having lower one)

⇒ we have a way of ordering objects in more than 1D

Another proposal:

2. Simplicial (or Liu) Depth [SD]

- The *simplicial depth* [Liu (1990)] at x w.r.t. F is defined to be

$$SD(F; x) = P_F\{x \in S[X_1, \ldots, X_{d+1}]\}$$ = ℙ(that x belongs to the simplex composed by a (d+1) dimensional sample)

- Here $S[X_1, \ldots, X_{d+1}]$ is a closed simplex formed by a sample of (d+1) random variables from F.
- The sample version of $SD(F; x)$. is obtained by replacing F by F_n, or alternatively, by computing the fraction of the sample random simplices containing the point x.

! Empirical counterpart ⟹

$$SD(F_n; x) = \binom{n}{d+1}^{-1} \sum_{\cdot} I_{(x \in S[X_{i_1}, \ldots, X_{i_{d+1}}])}$$

= proportion of the d+1 dimensional simplexes containing the point

where $I_{(\cdot)}$ is the indicator function.

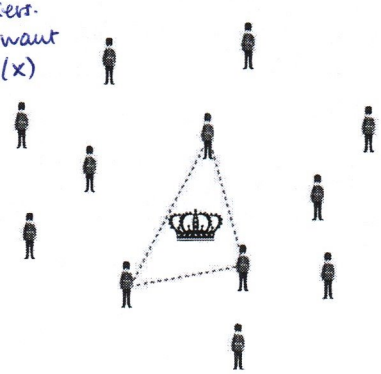
⇒ to compute the depth of each point we have to build up all the possible simplexes (containing that point) with all the (d+1) samples contained in the data cloud.

- In the bivariate case, the SD of a point $\mathbf{x}$ is the number of triangles with vertices in the data cloud S_n and containing $\mathbf{x}$.

Simplicial (or Liu) Depth [SD]

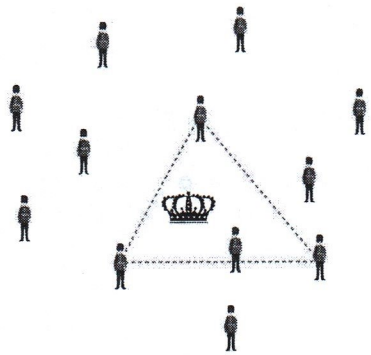
The points are the soldiers. The point for which we want to measure the depth (x) is the queen.

<u>Idea</u>: if we have to put the queen somewhere, we want to put her in the deepest point ("the most protected")

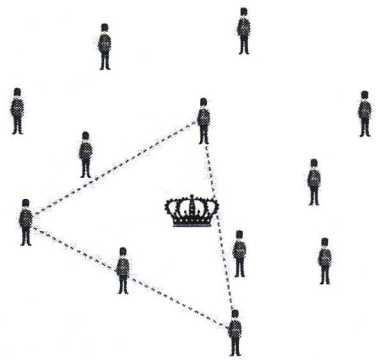


The simplices in $\mathbb{R}^2$ are triangles. To calculate the depth of x we build up all the possible triangles and we count in how many of them x is contained.

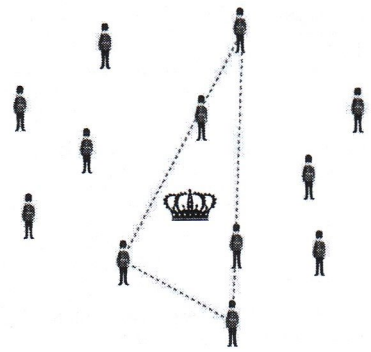
Simplicial (or Liu) Depth [SD]



Simplicial (or Liu) Depth [SD]

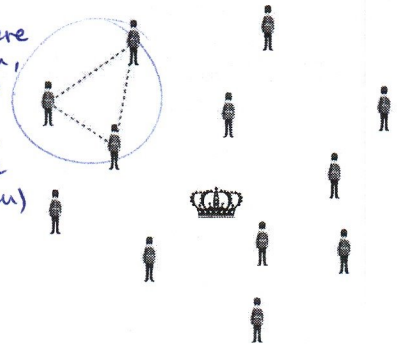


Simplicial (or Liu) Depth [SD]



Simplicial (or Liu) Depth [SD]

∃ triangles where the queen is not in, but these triangles are minimum in the number (w.r.t. all the other possible positions of the queen)



⟹ We got another method to calculate depths, however we are not gaining so much in terms of computation. (we still have to build all possible triangles for any point)

"similar" method:

3. Oja Depth [OD] (not so used)

- The *Oja depth* [Oja (1983)] at ***x*** w.r.t. *F* is defined to be

$$OD(F; x) = \left[1 + E_F\{\text{volume}(S[x, X_1, \ldots, X_d])\}\right]^{-1}$$

1 + expected value of the volume of the simplices that are build using **x** as one of the d+1 points composing the vertices

- Here $S[\mathbf{x}, X_1, \ldots, X_d]$ is the closed simplex with vertices **x** formed by a sample of d random variables from *F*.
- The sample version of $OD(F; \mathbf{x})$ is obtained by replacing *F* by F_n, or alternatively, by computing the fraction of the simplices of interest.

empirical version →

$$OD(F_n; x) = \binom{n}{d}^{-1}[1 + \sum\nolimits_{*}\{\text{volume}(S[x, X_{i_1}, \ldots, X_{i_d}])\}]^{-1}$$

where * indicates all d-plets $i_1, \ldots, i_d$ such that $1 \le i_1 \le \cdots \le i_d \le n$.

- *OD* of a point **x** is the sum of the volumes of every closed simplex having a vertex at **x** and the others in any *d* points of the data set.

---

4. Mahalanobis Depth [MHD]

- The *Mahalanobis Depth* [Mahalanobis(1936)] at ***x*** w.r.t. *F* is defined to be

$$MhD(F; x) = \left[1 + (x - \mu_F)\Sigma_F^{-1}(x - \mu_F)\right]^{-1}$$

the Mahalanobis depth is the Mahalanobis distance of the point from the other.

where μ_F and Σ_F stand for the mean vector and the dispersion matrix.

- The sample version can be obtained replacing μ_F and Σ_F with the corresponding sample estimates.
- This function fails at being robust, since it is based on measures of mean and variance, which are not *robust*.

⟹ this is why is not so common used

- Another disadvantage of this function is that it depends on the existence of second moments.

(∃ distributions that don't have finite second moment. For those we cannot evaluate Σ (variance-covariance matrix))

What this depth does?

Mahalanobis Depth [MHD]

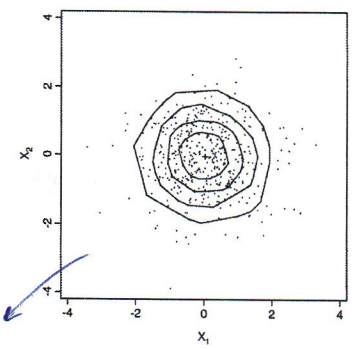


FIG. 3. *Normal contours by MD.*

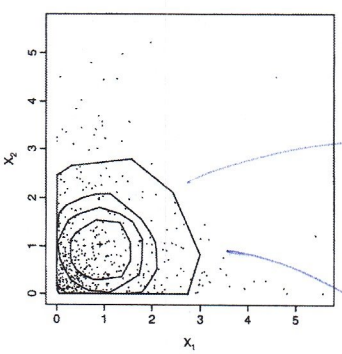


FIG. 4. *Exponential contours by MD.*

these are something similar to the level contours: all the points of the contour have the same depth

the Mahalanobi's depth contours are similar to the ones we know from gaussian, good!

here the level contours are not as good as with the gaussian

Why? Next slide

Remarks

- Note that as the Mahalanobis depth measures the quadratic distance from each sample point to the sample mean, the contours expand outward from the sample mean, following only the distance change and ignoring the asymmetric nature of the exponential sample.
- On the other hand, the simplicial depth is defined to measure the relative position of a point w.r.t. to a distribution and thus to capture the underlying probabilistic geometry.
- Similar explanations hold for the behavior of the half-space depth.

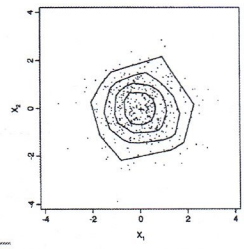


FIG. 1. *Normal contours by SD.*

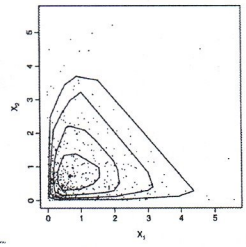
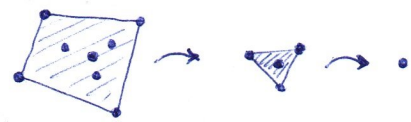


FIG. 2. *Exponential contours by SD.*

5. Convex Hull Peeling Depth [CD]

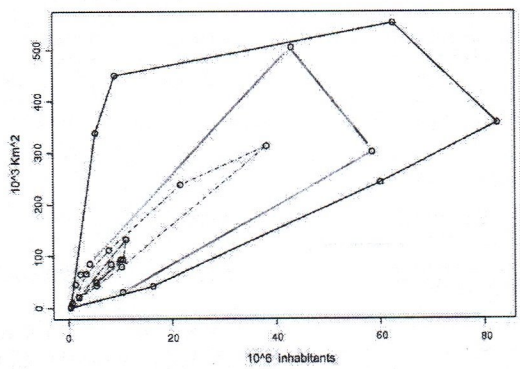
- The *Convex Hull Peeling Depth* [Barnett (1976)] at $\boldsymbol{x}$ w.r.t. the dataset $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ is simply the level of the convex layer $\boldsymbol{x}$ belongs to.
- A convex layer is defined as follows.
  - Construct the smallest convex hull which encloses all sample points $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$.
  - The sample points on the perimeter are designated the first convex layer and removed.
  - The convex hull of the remaining points is constructed; these points on the perimeter are the second convex layer.
  - The process is repeated, and a sequence of nested convex layers is formed.
- The higher layer a point belongs to, the deeper the point is within the data cloud.
- Despite appealing, the main drawback of this depth is the lack of an associated distribution theory.

given a cloud of data we can construct the convex hull. Then we remove the points that create the convex hull and we repeat the operation:



we obtain a sequence of nested convex layers, according to the layer each point belongs to we have a depth.

Example:



Convex hull graphic representing the total population and area of each of the 27 EU countries.

Further Depths da paper Liu

- Further definitions are possible:
  - Majority Depth (Singh, 1991)
  - Likelihood Depth (Fraiman and Meloche, 1999)
  - Spatial Depth (https://www.rdocumentation.org/packages/ddalpha/versions/1.3.11/topics/depth.spatial)
- Local depths ≠ global — it's used for local features
  - Paindaveine, D., Van Bever, G. (2013) From depth to local depth : a focus on centrality. Journal of the American Statistical Association 105, 1105–1119.
  - Kosiorowski D, Zawadzki Z (2020). *DepthProc An R Package for Robust Exploration of Multidimensional Economic Phenomena.*

Remarks

- Given a notion of data depth, one can compute the depths of all the sample points and order them according to decreasing depth values.
- This gives a ranking of the sample points from the center outward.
- Let $x_{[i]}$ denote the sample point associated with the *i*-th highest depth value. We view $x_{[1]}, \dots, x_{[n]}$ as the order statistics, with $x_{[1]}$ **being the *deepest*** or the *most central* point or simply the *center*, and $x_{[n]}$ **being the *most outlying*** point.

  => *larger rank is associated with a more outlying position w.r.t. the data cloud.*
- These order statistics induced by a data depth are different from the usual order statistics on the real line, since the latter are ordered from the smallest sample point to the largest, while the former start from the *middle* sample point and move outwards in all directions.
- The depth-induced order statistics are referred to as ***depth order statistics*** *(DO-statistics)*, and their ordering or ranking as *depth ordering* or ***depth ranking.***
- When ties occur in the ordering, the corresponding sample points are viewed as depth-equivalent, and the set of these points is termed a ***depth-equivalence class*** *(DE-class)* for short.

} however we can allow some noise for not having too many ties

Properties and Definitions

Let $D(x)$ be the depth of a point $x \in R^d$.

- **Definition - *level set* or *contour***

The set $\{x \in R^d : D(x) = t\}$ is called the *level set* or *contour* of depth t.

- **Definition**

The set $\{x \in R^d : D(x) > t\}$ is referred to as *region enclosed by the contour at depth p,* and denoted by R(t).

- **Definition – *p-th Central Region***

The set $C_p = \cap_t \{R(t) : P_F(R(t)) \geq p\}$ is referred to as the *p-th central region.*

region containing at least p% of the density of the distribution

In other words, C_p **is the smallest region enclosed by depth contours with a mass probability larger than or equal to *p***. The boundary of C_p is referred to as the *p-th level contour*, and is denoted by *Q(p)* or $Q_F(p)$ when we need to stress that *F* is the underlying distribution.

In theory, the empirical versions of C_p and $Q_F(p)$ should be the corresponding empirical versions. However, for computational and graphical convenience, the convex hull containing the most central fraction *p* sample points is used as the sample estimate of C_p.

Properties and Definitions

Denote $\wp$ the class of probability distributions on the Borel sets of R^d, and P_X the law of a given random vector $X \in R^d$.

Let $D(.,.) : R^d \times \wp \to R$ be a bounded nonnegative map such that the following properties hold:

1. ***Affine Invariance:*** for any dxd non singular matrix A and any vector b $\in R^d$

$$D(Ax + b, P_{Ax+b}) = D(x, P_x) \quad \forall x \in R^d$$

⇒ affine transformations of the data do not affect the ranking

2. ***Maximality at center:***

$$D(\theta, P_x) = \sup_{x \in R^d} D(x, P_x) \quad \forall P \in \wp \text{ centered at } \theta$$

⇒ the median is the deepest point

3. ***Monotonicity wrt the deepest point:*** for any P having deepest point at θ

$$D(x, P_x) < \quad D(\theta + \alpha(x - \theta), P_x) \quad \forall \alpha \in [0,1]$$

⇒ all the points that are not the deepest one have decreasing values of the depth (in this case: the larger the distance from θ, the "lower" the depth)

4. ***Vanishing at infinity:***

$$D(x, P_x) \to 0 \quad \text{as } ||x|| \to \infty$$

Then $D(.,P)$ is called a **Statistical Depth Function**.

The sample version of $D(.,P)$, denoted by $D(.,P_n)$, is defined by replacing *P* by the empirical version P_n. It can be proved to be consistent.

Apart from Oja and Likelihood Depths, all the others are affine invariant. Anyway, the deepest point and the order induced by all the measures is affine invariant.

All the depths (except **OJA**, which is not affine invariant) respect these properties.

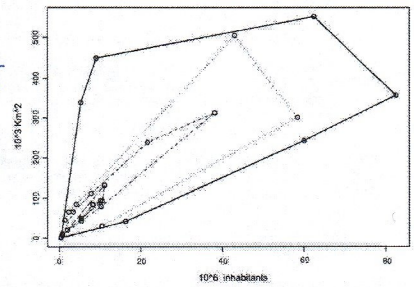
Graphical representation of Data Depth

- The center-outward ordering induced by a data depth may gives rise to a simple graphical techniques for presenting bivariate data sets.
- These techniques can be essentially viewed as a generalization of the univariate box-plot.
- Despite the concept of depth is pretty intuitive even in larger dimensions, the graphical representation of data depth is directly related to the difficulty of computing data depth functions. Even in the bivariate case it is a complex issue and sometimes it is still under development.
- The most common **graphical tools for visualizing the depths** o a dataset are:

1. The convex hull
2. The Bagplot
3. The Sunburst plot
4. The DD-plot

1. Convex Hull

- The convex hull graphic is based on convex hull peeling depth.
- The central idea of convex hull is to construct **convex layers that enclose all data set points**.
- Advantages: extremely intuitive -> A point lying on the most external layer is obviously less deep than one lying on an internal layer.
- Disadvantages: based on convex hull peeling depth, which may not be always the most appropriated



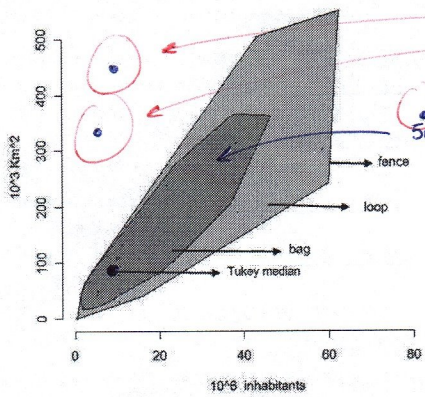
Convex hull graphic representing the total population and area of each of the 27 EU countries.

2. Bagplot

- The **bagplot** was proposed by Rousseeuw et al. (1999) and can be considered a **generalization of the univariate boxplot.**
- The main components of the bagplot are:
  - The **bag**, containing **50% of the observations (deepest 50%)**.
  - Within the bag is the **Tukey median**, the observation with the maximal depth.
  - The **fence**, which **separates the outliers from the other observations**.
  - A **loop**, where lie the observations that do not belong to the bag but are inside the fence.
- The bagplot shows several characteristics of the data:
  - ***Location*** (the depth median),
  - ***Spread*** (the size of the bag),
  - ***Correlation*** (the orientation of the bag),
  - ***Skewness*** (the shape of the bag and the loop),
  - ***Tails*** (the points near the boundary of the loop and the ouliers).

Bagplot

Bagplot graphic representing the total population and area of each of the 27 EU countries



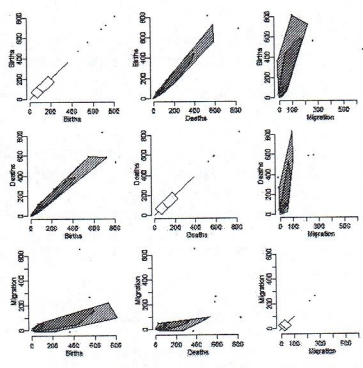
outliers (since ∉ loop)

50% of the observations, the deepest

What about higher dimensions? Bagplot is effective up to 3 dimensions.

Bagplot

- The case presented above is bivariate. However, the location depth and the Tukey median may be considered in any dimension, being possible to define the bag in the p-dimensional case.
- In three dimensions the bag is a convex polyhedron, but in higher dimensions it becomes difficult to visualize it.
- One option to the p-dimensional case, is to represent the depth in a **bagplot matrix**, i.e. a matrix containing the bagplot for each pair of two variables.
  The **diagonal** of the matrix is the boxplot of each variable, since the bagplot reduced to the unidimensional case is the boxplot.



Matrixbagplot graphic representing the number of births, the number of deaths and the migration of each of the 27 EU countries

We can actually build a 3D bagplot or we can split the 3D bagplot and make a bagplot for every couple of variables:

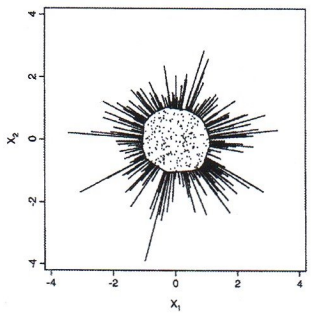
(on the diagonal we have the boxplot for each variable)

3. Sunburst plot

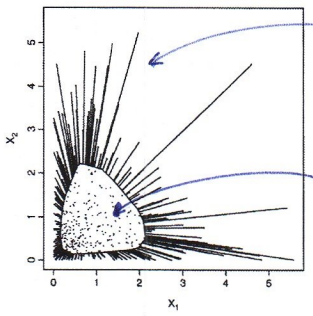
analog to the bagplot

- Analogous to Bagplot
- The plot resembles the sun with its rays radiating in all directions.
- For a given sample,
  - Apply a data depth to identify its center and the central 50% sample points.
  - Mark the center and draw the contour to enclose the 50% central hull.
  - The rays in the plot are obtained by joining the sample points outside of the 50% central hull to the center, keeping only the segments outside the contour.
- The center, the central region and the rays obtained this way can be regarded as the analogues of the median, the interquartile range and the whiskers in the box-plot.
- The sunburst plot and the contours plot provide a quick and informative overview of the **shape, concentration, spread** and **skewness** of the underlying distribution for a given sample.

Sunburst plot



Sunburst plot for a Gaussian sample



Sunburst plot for an Exponential sample

all the other data (not in the 50% deepest) are connected to the deepest point but we see only the out-of-bag connection

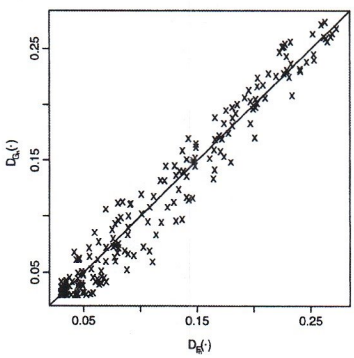
the central bag contains 50% of the data, the more in deep ones.

4. DD plot

Useful for comparing distributions of different samples

(we could do the QQ-plot, but it's different: in the QQ-plot we compare the empirical quantiles with the theoretical ones of a given distribution; how we want to understand if two distributions are similar, we don't mind which kind of distributions)

- Let **x** and **y** be two n- and m-dimensional samples from *F* and *G*, respectively.
- The *DD-plot* plots the depth values of the combined n+m dimensional sample (**z**) under the two corresponding empirical distributions F_n and G_m.
- If the two given distributions are identical, then these plots are segments of the diagonal line in R^2.
- Plots which **deviate from this straight line** indicate differences between the two distributions.
- Different deviation patterns in the plot correspond to different types of variations between the distributions, for example **location shifts** or scale increases. Analogous to QQ plot



We have samples from F_n and G_m. We take all the samples together and call them z_i (so they're indistinguishable). For every point we compute the depth w.r.t. F and w.r.t. G. For every point we have 2 depths. We plot these points and if they're ~ on a straight line we can say that F and G are the same!
(because the points have nearly the same depth according to both F and G ⇒ points coming from F or G are the same)

Descriptives & Estimation

- Given a method for ordering objects of a multivariate sample, a rank-based estimation theory may be developed.
- In particular, traditional univariate measures may be translated in the corresponding indexes in the multivariate case, for describing and comparing distributions.
  - Location (mean (e.g.))
  - Dispersion (matrix or scalar) (variance (e.g.))
  - **Shape: Skeweness (spherical, elliptical, antipodal, angular symmetry) and Kurtosis**

! *Note that information like location, dispersion and shape indexes are extremely useful when dealing with multivariate distributions with no parametric assumption holding.*

(since we have no other information about distributions)

Robust vs Nonparametric estimation

- There is often a misleading understanding about the relationship between nonparametric and robust statistics.

They are, indeed, connected, but definitively NOT the same!

- **Robust statistics** seek to provide methods that emulate popular statistical methods, but which are **not unduly affected by outliers or other departures from model assumptions**.
- In statistics, classical estimation methods rely heavily on assumptions which are often not met in practice.
- Unfortunately, when there are outliers in the data, **classical estimators often have very poor performance**, when judged using the *breakdown point* and the *influence function*.

We want robustness because we know nothing about the distributions of the data and we don't want our inference to be influenced by outliers (moreover we want to work in multidimensional settings, and as we grow with dimension, it grows the probability of getting outliers as well)

Robust vs Nonparametric estimation

- Robust statistics aims at providing estimators (of location, dispersion, shape, ...) which are **not (less) affected by the presence of outliers or deviations from model assumptions**.
- Such estimators are related, but not identical to non-parametric statistics, where we drop the hypothesis of underlying parametric distribution. The main reason of overlapping is the fact that methods based on ranks heavily enter the play.
- The main approaches adopted in robust statistics are:
  - ***Discard data*** (that make statistics not robust)
  - ***L-estimators***: Use linear combinations of order statistics
  - ***R-estimators***: Use rank instead of values
  - ***M-estimators***: Estimators based on Maximum-likelihood argument
- A nice introduction on these topics may be found at http://www.isdc.unige.ch/~paltani/Courses/robust_paltani.pdf

---

How can we measure robustness?

- Robust statistics aims at providing estimators (of location, dispersion, shape, ...) which are **not (less) affected by the presence of outliers or deviations from model assumptions**.

*How can we **measure robustness** of an estimator?*

- The basic **tools used to describe and measure robustness** are
  - the ***breakdown point***,
  - the ***influence function***
  - the ***sensitivity curve*** *(for parametric case, not of interest here)*.

---

1. Breakdown Point of an Estimator

= percentage of points needed to make the estimator "corrupted" (= different)

- In the development of robust estimators of statistical parameters, a key quantity in measuring their performances is the breakdown point.
- Given a dataset **X** and a real value estimator T(**X**), its breakdown point is:

$$\varepsilon^* = \inf\left\{\varepsilon : \sup_{\mathcal{X}'_\varepsilon} |T(\mathcal{X}) - T(\mathcal{X}'_\varepsilon)| = \infty\right\}$$

$\mathcal{X}'_\varepsilon$: Contaminated/modified version of dataset X

proportion of data that we're using for corrupting the original database

=> the breakdown point measures the minimum data corruption required to bring estimates very far away from the correct values.

- Sample median breakdown point = 0.5
  (the highest value a translation-equivariant estimator can achieve)
- Sample Mean breakdown point = 0
  (just one observation can cause breakdown)

Note: This notion makes no assumption on the particular functional form of the dataset corruption considered.

For example: in the case of the mean it's enough one point to have a drastic change. The median does not move much because we added just one data. In order to move the value of the median we need to move at least half of the data ($\Rightarrow$ the breakdown of the data is 1/2 (higher breakdown point achievable))

2. Empirical Influence Function (EIF)

- The Empirical Influence Function (EIF) is a measure of the dependence of the estimator on the value of one of the points in the sample.
- It is a model-free measure in the sense that it relies on calculating the estimator again with a different sample.
- **Idea: replace the *i*-th value in the sample by an arbitrary value and look at the output of the estimator.**

Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\{X_1, \dots, X_n\}$ be a sample and T(X) an estimator. Let $i \in \{1, \dots, n\}$. The EIF_i at observation i is defined as

$$EIF_i : x \in \mathcal{X} \mapsto n \cdot (T_n(x_1, \dots, x_{i-1}, x, x_{i+1}, \dots, x_n) - T_n(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n))$$

- **Note:** Useful for testing reliability and robustness of prediction for any model, and especially for black-box models (e.g., Machine Learning)
- **Note:** Alternatively, the EIF is defined scaling by n+1 instead of n the effect on the estimator of adding the point x to the sample

Location

- We would like now to use depths as robust estimators of distribution features.
- Given a notion of data depth, there is a natural choice of **location parameter** for the underlying distribution, namely the deepest point or the average of the deepest points if there is more than one.
- For the same distribution, different notions of depth may lead to **different deepest points**, which may be quite different **if the underlying distribution is asymmetric**.
- Viewing **the deepest point as the sample median**, we may state the following general property regarding its *distributional symmetry and unbiasedness* (valid also in the case of no-moment distribution, e.g. the Cauchy):
  - **PROPOSITION**: *If the population distribution is symmetric, then the distribution of any affine invariant sample median (the deepest point) is also symmetric about the population center of symmetry.*
- => the half-space median, simplicial median and all deepest points derived from the depths listed before are **unbiased estimators for the mean** of a multivariate Normal distribution.

http://cgm.cs.mcgill.ca/~athens/Geometric-Estimators/location.html

why the median? If d ($\mathbb{R}^d$) increases it really becomes difficult to characterize the joint distribution of the data, but the more important thing is trying to find outliers (which can give a lot of infos.). The median is useful in this task: median is less sensitive (⇒ more robust) w.r.t. the mean.

Scale or Dispersion

- There are two common **approaches to quantify the dispersion of a multivariate distribution**: as a **matrix** or as a **scalar**.

(we can choose the form)

Matrix form of scale or dispersion

$$\mathbf{S}_n(t) = \begin{cases} (X_{[i]} - \nu_n)(X_{[i]} - \nu_n)', & \text{for } \frac{i-1}{n} < t \le \frac{i}{n}, \\ \mathbf{O}, & \text{for } t = 0. \end{cases}$$

median

it's like the variance-covariance matrix but with the median instead of the mean (we can interpret this object as variance-covariance correlator)

where ν_n is the deepest sample point, and **O** is the zero matrix.

- We can interpret loosely the entries of $\mathbf{S}_n$ as variations and covariations, without requiring the existence of the corresponding moments.
- We can also imitate the definition of the so-called generalized sample variance in classical multivariate analysis by taking the determinant of our sample scale matrix $\mathbf{S}_n$ and using the resulting *single* numerical value to describe the variation expressed by $\mathbf{S}_n$.
- This determinant will be called the *generalized sample scale*.

we have the matrix form and we can synthetize the information with one scalar: the determinant

or

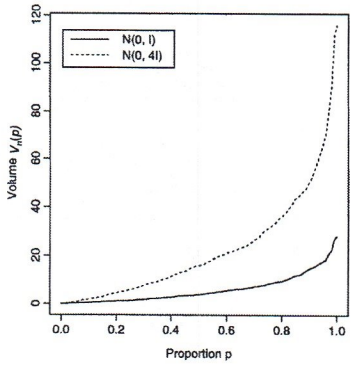
Scale or dispersion

Scalar form of scale dispersion

C_p = region having p% of the density of the distribution

- A different measure of scale or dispersion of a distribution can be defined by keeping track of **how the *p*-th central region C_p expands as *p* increases**. This is a distributional property which can be characterized easily by the speed with which the data depth decreases.
- Recall that, for a given data depth $D(.,.)$, the level sets $\{x: D(x, F) = c\}$ form nested contours as the level *c* decreases. Thus we can define a scale curve by taking the plot of *p* versus $V_n(p)$, where

$$V_n(p) = \text{volume}\{C_p\}$$



how the dispersion of the distribution evolves

p% contained density vs. volume of the C_p region

The dashed line grows more (reasonable!)

Scale or Dispersion

- In the Figure above, C_p is the p-th central region.
  The sample scale curve, **$V_n(p)$, is simply the volume of the convex hull containing $\lceil np \rceil$ most central points.**
- Clearly, the **faster growing** $V_n(p)$ is associated with a **larger dispersion** of the distribution.
- C_p is the central region amassing probability p, which expands as p grows.
- For $p_1 < p_2$, the difference $V(p_2) - V(p_1)$ reflects the central probability increment speed relative to the central-region expansion from C_{p1} to C_{p2}.
- In other words, if the scale curve of G is consistently above the scale curve of F, then G has a larger dispersion than F.

  → *Nonparametric way to determine stochastic dominance among distributions*

Modeling, Prediction and possible extensions

- **Regression Depths & Deepest Regression**
  - Rousseeuw and Hubert (1999).
  - Van Aelst et al (2000)
- **Depths for Functional Data** *(see the next presentation)*
  - Ieva, F., Paganoni, A.M., Romo, J., Tarabelloni, N. (2019) roahd Package: Robust Analysis of High Dimensional Data *R Journal.*
  - Tarabelloni, N., Ieva, F., Paganoni, A.M., Biasi, R. (2015) Use of depth measure for multivariate functional data in disease prediction: an application to electrocardiograph signals. International Journal of Biostatistics, 11(2), 189–201
  - Ieva, F., Paganoni, A.M. (2015) Discussion of "multivariate functional outlier detection" by M. Hubert, P. Rousseeuw and P. Segaert. Statistical Methods and Applications, 24 (2): 217-221
  - Ieva, F., Paganoni, A.M.. (2013) Depth Measures for Multivariate Functional Data. Communication in Statistics – Theory and Methods, 42(7): 1265-1276.
- **Depth-based Control Charts**
  - Barale, M. S., Shirke, D. T. (2019) Nonparametric Control Charts Based on Data Depth for Location Parameter. Journal of Statistical Theory and Practice, volume 13, Article number: 41 (2019)
  - Cascos, I., López-Díazb, M. (2018) Control charts based on parameter depths. Applied Mathematical Modelling, Volume 53, January 2018, Pages 487-509 https://doi.org/10.1016/j.apm.2017.09.009

Take home messages

- Nonparametrics is not a new approach, but it is becoming more and more useful and adopted as long as the data dimensionality and complexity grows up.
- Depth measures are a powerful descriptive and inferential tool in nonparametric multivariate statistics.
- The development of methods and algorithms is an active and recent field of research.
- Computational issue is an issue!
- Robust statistics and Nonparametrics are not the same, but they are related.

References

- Barnett, V. (1976). The ordering of multivariate data (with discussion). Journal of the Royal Statistical Society. Series A. (General), 139: 318–352.
- Brown, B.M.; Hettmansperger, T.P. (1989). An Affine Invariant Bivariate Version of the Sign Test. Journal of the Royal Statistical Society: Series B (Methodological), 51:117-125.
- Christmann, A. (2002). Classification based on the SVM and on regression depth. In Dodge, Y. (eds),Statistical data analysis based on the L1 norm and related methods, Basel: Birkhouser, 341-352.
- Claeskens, G.; Hubert, M.; Slaets, L.; Vakili, K. (2014). Multivariate Functional Halfspace Depth. Journal of the American Statistical Association, 109 (505):411-423.
- Donoho, D.L.; Gasko, M. (1992). Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness. Annals of Statistics, 20(4):1803-1827.
- Fraiman, R.; Liu, R; Meloche, J. (1997). Multivariate density estimation by probing depth. In L1-Statistical Procedures and Related Topics. Hayward, CA: IMS, 415-430.
- Fraiman, R. and J. Meloche (1999). Multivariate L-estimation. Test, 8: 255–317.
- Genest, M., Masse, J.C., Plante, J.F. (2012). depth: Depth functions tools for multivariate analysis, http://CRAN.R-project.org/package=depth
- Hettmansperger, T., Nyblom, J. and Oja, H. (1992). On multivariate notions of sign and rank. In Dodge,Y. (editores), L1 Statistics and Related Methods. Amsterdam: Elsevier, 267-278.
- Hettmansperger, T.P.; Oja, H. (1994) Affine Invariant Multivariate Multisample Sign Tests. Journal of the Royal Statistical Society: Series B (Methodological), 56 (1):235-249.
- Ieva, F.; Paganoni, A.M. (2020). Component-wise outlier detection methods for robustifying multivariate functional samples. Statistical Papers, 61: 595–614.
- Liu, R.Y. (1990) On a Notion of Data Depth Based on Random Simplices. Annals of Statistics, 18 (1):405-414.
- Jörnsten, R. (2004). Clustering and classification based on the L1 data depth. Journal of Multivariate Analysis, 90 (1):67-89.

References

- Liu, R.Y.; Parelius, J.M.; Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference (with discussion). Annals of Statistics, 27:783–858.
- López-Pintado, S., & Romo, J. (2009). On the concept of depth for functional data. Journal of the American Statistical Association, 104:718–734.
- Mahalanobis, P. C. (1936). On the generalized distance in statistics. Proceedings of the National Acadedmy of Science of India,12: 49-55.
- Oja, H. (1983). Descriptive statistics for multivariate distributions. Statistics and Probability Letters,1,327-332.
- Rousseeuw, P., Ruts, I. and Tukey, J. (1999). The Bagplot: A bivariate boxplot. The American Statistician,53 (4):382-387.
- Ruts, I; Rousseeuw, P. (1996). Computing depth contours of bivariate point clouds. Computational Statistics & Data Analysis 23:153-168.
- Serfling, R. (2004). Nonparametric multivariate descriptive measures based on spatial quantiles. Journal of Statistical Planning and Inference,123:259-278.
- Serfling, R. (2006). Depth functions in nonparametric multivariate inference. DIMACS Series in Discrete Mathematics and Theoretical Computer Science,72:1-16.
- Sun, Y.; Genton, M. G. (2011) Functional boxplots, Journal of Computational and Graphical Statistics 20: 316–334.
- Tarabelloni, N.; Ieva, F.; Paganoni, A.M.; Biasi, R. (2015). Use of depth measure for multivariate functional data in disease prediction: an application to electrocardiograph signals. International Journal of Biostatistics, 11 (2):189–201.
- Tukey, J. (1975). Mathematics and the picturing of data. In Proc. 1975 Inter. Cong. Math., Vancouver 523–531. Montreal: Canad. Math. Congress.
- Van Aelst, S., Van Driessen, K. and Rousseeuw, P. (2000). A robust method for multivariate regression. InKiers, H. A. L. (eds), Data analysis, classification, and related methods. Berlin: Springer, 309-315.

References

- Wang, J. and Serfling, R. (2005). Nonparametric multivariate kurtosis and tailweight measures. Journal of Nonparametric Statistics,17:441-456.
- Zhang, J. (2002). Some extensions of Tukey's depth function. Journal of Multivariate Analysis, 82:134-165.
- Zuo, Y. amd Serfling, R. (2000), General Notions of Statistical Depth Functions. Annals of Statistics, 28 (2):461–482.

Packages

- **roahd** – Ieva F, Paganoni AM, Romo J, Tarabelloni N (2019). "roahd Package: Robust Analysis of High Dimensional Data." *The R Journal*, **11**(2), 291–307. https://doi.org/10.32614/RJ-2019-032.
- **depth** - Genest, Masse, Plante, (2012). depth: Depth functions tools for multivariate analysis
- **DepthProc** - Kosiorowski D, Zawadzki Z (2020). *DepthProc An R Package for Robust Exploration of Multidimensional Economic Phenomena*.
- **aplpack** – Wolf, Bielefeld (2014) aplpack: Another Plot PACKage: stem.leaf, bagplot, faces, spin3R, plotsummary, plothulls, and some slider functions
- **fda** - Ramsay, Wickham, Graves, Hooker, (2013) fda: Functional Data Analysis