The Concept of Depth In Statistics

Maria Raquel Neto

Abstract:

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 non-parametric methods to be used in multivariate data analysis.

Different depth measures with different characteristics are presented in this work. All these measures have been proposed with the same objective, to determine the data depth of an observation.

Other challenges within the topic of data depth are the computational implementation of the depth measures and the graphical representation of the data depth.

The field of applications of data depth is vast and is still growing. Throughout this paper references are made to some of these applications, one of the most interesting being the deepest regression, a robust linear regression method. The deepest regression is based on the regression depth, a function that allows the calculation of the depth of a regression line.

The ideas and results around the notion of data depth are complex but very interesting and useful. It is expected that the studies currently under way will produce the developments that will make their practical implementation easier than they are today. Here only simple but fundamental aspects are discussed.

Keywords: Depth, rank, robustness, deepest regression.

1. Introduction

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.

Since 1975, when Tukey (1975) introduced the location depth, many more depth functions have been proposed. Some of these functions are defined in Section 2. Most depth functions are robust and affine invariant making them well suited for the study of real life high dimensional data sets that may contain outliers.

The concept of data depth has many applications in the multivariate analysis field. Some of this applications are nonparametric description of multivariate distributions (see Liu et al.,1999; Serfling, 2004; and Wang and Serfling, 2005), outlier identification (see Serfling, 2006; and Zhang, 2002), depth-based classification and clustering (see Ruts and Rousseeuw, 1996; Christmann, 2002; and Jörnsten, 2004), rank and sign tests (see Brown and Hettmansperger, 1989; Hettmansperger et al., 1992; and Hettmansperger and Oja, 1994), multivariate density estimation (see Fraiman et al., 1997) and data based linear regression (see Rousseeuw and Hubert, 1999).

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.

Also the graphical representation of data depth is an important issue to bear in mind. In Section 3 some data depth graphical representations are introduced, such as the bagplot and the convex hull peeling graphics.

Rousseeuw and Hubert (1999) extended the notion of depth to the linear regression, introducing the regression depth. This depth measures how well a hyperplane fit represents the data and is the basis to a new regression estimator, the deepest regression estimator. Regression depth and deepest regression estimator are introduced in Section 4.

In Section 5 some similarities between location depth and regression depth are discussed.

2. Depth functions

In 1975, Tukey (1975) introduced the concept of location depth (also called halfspace depth and Tukey depth). In the bivariate case, the location depth of a point $\boldsymbol{x}$ relative to a bidimensional data set S_n is defined as the smallest number of data points lying in one of the sides of a line passing through $\boldsymbol{x}$. This definition can be extended to higher dimensions:

Definition 1 - Location Depth (Tukey, 1975)
Location depth of a point $\boldsymbol{x} = (x_1, ..., x_p) \in S_n = \{\boldsymbol{x_i} = (x_{i1}, ..., x_{ip}); i = 1, ..., n\} \subset \mathbb{R}^p$ relative to a p-dimensional data set S_n is defined as the smallest number of data points in a closed halfspace with boundary through $\boldsymbol{x}$.

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. 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.

There are many depth functions, but they all have the same purpose, to measure how deep (or central) a given point $\boldsymbol{x}$ is. Some useful examples of other data depth functions are:

Definition 2 - Mahalanobis Depth (Mahalanobis, 1936)
Mahalanobis depth of a point $\boldsymbol{x} \in S_n \subset \mathbb{R}^p$ relative to a p-dimensional data set S_n is defined as:

$$\mathrm{M}_h\mathrm{D}(\boldsymbol{x}; S_n) = \left[1 + (\boldsymbol{x} - \bar{\boldsymbol{x}})^T S^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}})\right]^{-1}$$

where $\bar{\boldsymbol{x}}$ and S are the mean vector and dispersion matrix of S_n.

This function fails at being robust, since it is based on non robust measures such as the mean and the dispersion matrix. Another disadvantage of this function is that it depends on the existence of second moments.

Definition 3 - Convex Hull Peeling Depth (Barnett, 1976)
Convex hull peeling depth of a point $\boldsymbol{x} \in S_n \subset \mathbb{R}^p$ relative to a p-dimensional data set S_n is simply the level of the convex layer to which $\boldsymbol{x}$ belongs to.

A convex layer is defined as follows. Construct the smallest convex hull which encloses all data set points. The 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. The disadvantages of convex hull peeling depth are: it is not a robust measure and is impossible to associate it a theoretical distribution.

Definition 4 - Oja Depth (Oja, 1983)
Oja depth of a point $\boldsymbol{x} \in S_n \subset \mathbb{R}^p$ relative to a p-dimensional data set S_n is defined as the sum of the volumes of every closed *simplex* having a vertex at $\boldsymbol{x}$ and the others in any p points of the S_n data set.

In the bivariate case, the Oja depth of a point $\boldsymbol{x}$ relative to a bivariate data set S_n is the sum of the areas of all triangles whose vertices are $\boldsymbol{x}, \boldsymbol{x_i}, \boldsymbol{x_k}$, with $\boldsymbol{x_i}$ and $\boldsymbol{x_k}$ belonging to S_n.

Definition 5 - Simplicial Depth (Liu, 1990)
Simplicial depth of a point $\boldsymbol{x} \in S_n \subset \mathbb{R}^p$ relative to a p-dimensional data set S_n is defined as the number of closed *simplex* containing $\boldsymbol{x}$ and having p+1 vertices in S_n.

In the bivariate case, the simplicial depth of a point $\boldsymbol{x}$ is the number of triangles with vertices in S_n and containing $\boldsymbol{x}$.

Does not exist one depth function that is always better than the others. In some cases, one function may be better because it fits the data set better or just because it is easier to calculate. However, there are some desirable properties that all data depth functions should ideally satisfy such as, affine invariance, maximality at the center, monotonicity relative to deepest point and vanishing at infinity (see Serfling and Zuo, 2000).

3. Graphical representation of data depth

The graphical representation of data depth is directly related to the difficult of computing data depth functions. Even in the bivariate case it is a complex issue and is still under development.

The most common ways of visualizing data depth are based on the graphical techniques of convex hull and bagplot. These techniques are based on different data depth functions, respectively on convex hull peeling depth and on location depth, producing some times identical results. Both these graphical techniques have the same purpose, to represent data depth and consequently the ranking of multivariate data.

3.1. 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. The process of constructing the convex layers was already explained in section 2.

A great advantage of this type of representation is the fact that it is extremely intuitive. A point lying on the most external layer is obviously less deep than one lying on an internal layer. A disadvantage is that it uses only the convex hull peeling depth, which may not be always the most appropriated.

In Figure 1 an example of a bivariate convex hull is presented.

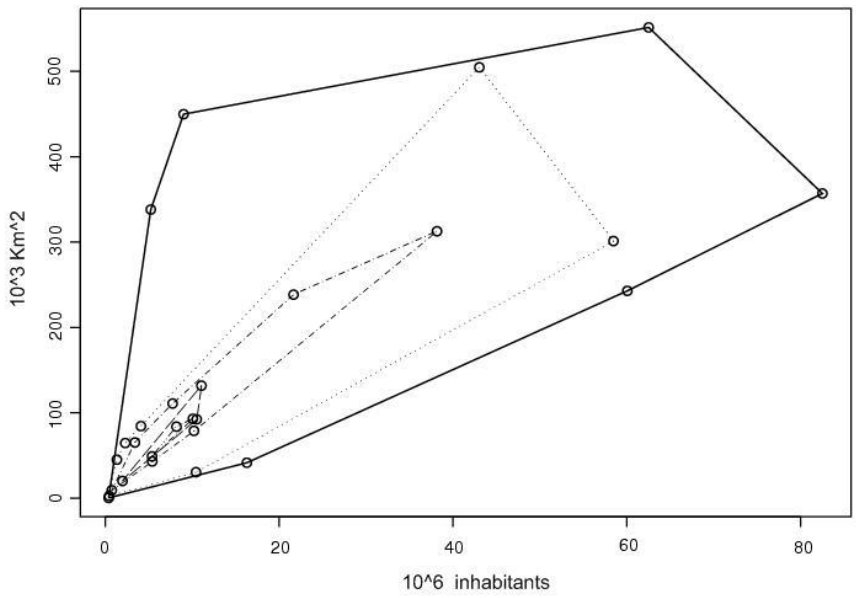

Figure 1: Convex hull graphic representing the total population and area of each of the 27 EU countries.

3.2. Bagplot

The bagplot was proposed by Rousseeuw et al. (1999) and can be considered a generalization of the famous univariate boxplot. This data depth representation is constructed using the location depth measure.

The bagplot is basically a scatterplot. The main component of the bagplot is the bag which contains 50% of the observations, and within the bag is the Tukey median, the observation with the maximal depth. This graphic is also composed by a fence, which separates the outliers from the other observations, and a loop, where lie the observations, that do not belong to the bag but are inside the fence.

Like the univariate boxplot, the bagplot shows several characteristics of the data: its 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), and tails (the points near the boundary of the loop and the ouliers).

An example of a bagplot with its various components can be seen in Figure 2.

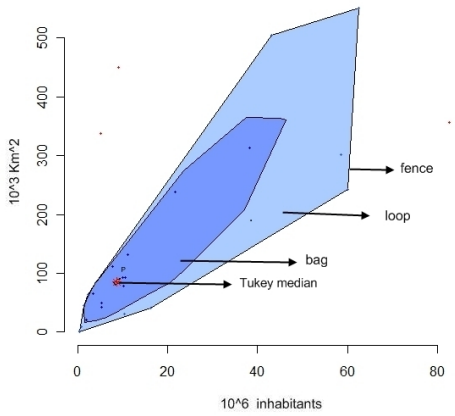

Figure 2: Bagplot graphic representing the total population and area of each of the 27 EU countries.

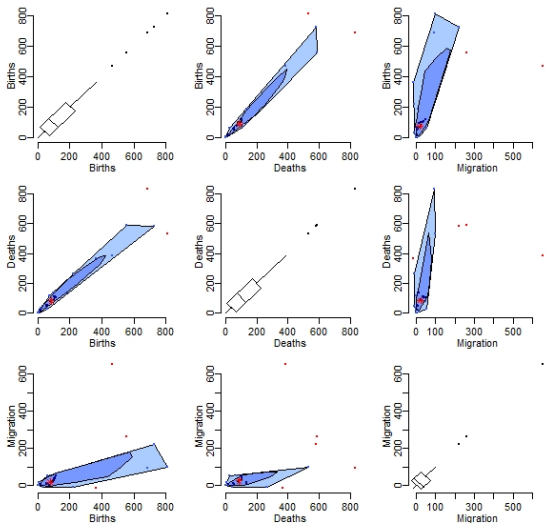

Figure 3: Matrixbagplot graphic representing the number of births, the number of deaths and the migration of each of the 27 EU countries.

The example presented in Figure 2 is the bivariate case. 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 combination 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. An example of the matrix bagplot can be seen in Figure 3.

4. Regression depth and deepest regression

In this Section the notions of regression depth and deepest regression are introduced. Regression depth provides the rank of any line (plane), rather than ranks of observations. The notion of regression depth forms the basis of the deepest regression estimator, which will be explained later on. First regression depth is introduced in the simplest case, the simple regression, and then generalized to the multiple regression case. Some more results about regression depth can be seen in Rousseeuw and Hubert (1999).

4.1. Depth in simple regression

In simple regression the goal is to fit a straight line $y = \beta_0 + \beta_1 x$ to a data set $Z_n = \{(x_i, y_i); i = 1, ..., n\} \subset \mathbb{R}^2$. All candidate fits will be denoted as $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^T$, where $\hat{\beta}_0$ is the intercept estimate and $\hat{\beta}_1$ is the slope estimate. The residuals are then denoted as $r_i = r_i(\hat{\boldsymbol{\beta}}) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. To simplify the notation, from here on $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ will be used instead of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^T$ and, consequently, $r_i(\boldsymbol{\beta})$ instead of $r_i(\hat{\boldsymbol{\beta}})$.

The regression depth of a candidate fit $\boldsymbol{\beta}$ indicates how well $\boldsymbol{\beta}$ fits the data. To define regression depth it is necessary to first define what is considered a nonfit.

Definition 6 (Rousseeuw and Hubert, 1999)
A candidate fit $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ to Z_n will be called a nonfit iff there exists a real number $v_\beta = v$ which does not coincide with any x_i and such that

$$r_i(\boldsymbol{\beta}) < 0 \;\; \forall x_i < v \quad \text{and} \quad r_i(\boldsymbol{\beta}) > 0 \;\; \forall x_i > v \qquad \text{or} \qquad r_i(\boldsymbol{\beta}) > 0 \;\; \forall x_i < v \quad \text{and} \quad r_i(\boldsymbol{\beta}) < 0 \;\; \forall x_i > v.$$

The existence of the real number v corresponds to the presence of a tilting point around which the line can be rotated until it is vertical, while not passing through any observation. Note that a line lying above or below all the observations is always a nonfit.

Now regression depth can be defined.

Definition 7 - Regression depth (Rousseeuw and Hubert, 1999)
The regression depth of a fit $\boldsymbol{\beta}$ relative to a data set Z_n is the smallest number of observations that need to be removed to make $\boldsymbol{\beta}$ a nonfit. Equivalently, $\text{rdepth}(\boldsymbol{\beta}, Z_n)$ is the smallest number of residuals that need to change sign to make $\boldsymbol{\beta}$ a nonfit.

From Definition 7, is easy to see that the limits for regression depth are $0 \leq rdepth(\boldsymbol{\beta}, Z_n) \leq n$, i.e. regression depth is at most n and at least 0. The maximum value of regression depth is achieved when all observations lie on a line. Some more results on the limits of regression depth can be seen in Rousseeuw and Hubert (1999).

Note that Definitions 6 and 7 do not require any distributional assumptions.

4.2. Depth in multiple regression

In multiple regression the goal is to fit an affine hyperplane $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{p-1})^T$ to a data set $Z_n = \{(x_{i1}, ..., x_{i,p-1}, y_i)\,; i = 1, ..., n\} \subset \mathbb{R}^p$. The x-part of each data point will be denoted by $\boldsymbol{x_i}^T = (x_{i1}, ..., x_{i,p-1}) \in \mathbb{R}^{p-1}$.

Definition 7 of regression depth is valid for multiple regression, but what is a nonfit in multiple regression has to be defined.

Definition 8 (Rousseeuw and Hubert, 1999)
A fit $\boldsymbol{\beta} = (\beta_0, ..., \beta_{p-1})^T$ is called a nonfit to Z_n iff there exists an affine hyperplane V in x-space such that no x_i belongs to V, and such that $r_i(\boldsymbol{\beta}) > 0 \;\; \forall x_i$ in one of its open halfspaces, and $r_i(\boldsymbol{\beta}) < 0 \;\; \forall x_i$ in the other open halfspace.

The above definition is equivalent to saying that a hyperplane is considered a nonfit when is possible to rotate it to the vertical position, without passing through any observation. This definition is similar to the one given for simple regression, the difference being that in simple regression the nonfit is always a line, which is much easier to visualize. In multiple regression, the visualization of a nonfit is still possible for $p = 3$, where the nonfit is a plane, but it becomes almost impossible to visualize in higher dimensions.

From the regression depth distribution it is possible to derive tests and confidence regions for the true unknown parameters in the linear regression model (see Van Aelst *et al*, 2000).

4.2. Deepest regression

The regression depth measures the quality of any candidate fit. Fits with higher regression depth fit the data better than do fits with lower regression depth. Hence, the regression depth ranks all possible fits from worst (rdepth=0) to best (maximal depth). This immediately leads to the definition of the deepest regression estimator.

Definition 9 - Deepest regression estimator (Rousseeuw and Hubert, 1999)
The deepest regression estimator $DR(Z_n)$ is the fit $\boldsymbol{\beta}$ with maximal regression depth relative to the data set, i.e.,

$$DR(Z_n) = \text{argmax}_{\boldsymbol{\beta}}\; rdepth(\boldsymbol{\beta}, Z_n)$$

In the univariate case it is easy to see that the deepest regression of a data set is its median. Hence, deepest

regression generalizes the univariate median to linear regression.

Some advantages of this method of regression are: i) it is highly robust against outliers (for more results about deepest regression robustness see Van Aelst and Rousseeuw, 2000); ii) it does not require any distributional assumptions. The deepest regression method can be applied to the usual regression models but can also be applied to more general regression functions (see Van Aelst et al., 2000).

5. Similarities between location depth and regression depth

Rousseeuw and Hubert (1999) proposed a unifying concept of depth, from which both location depth and regression depth can be derived. They define depth$(\boldsymbol{\beta}, Z_n)$ as the smallest number of observations of Z_n that would need to be removed in order to make $\boldsymbol{\beta}$ a nonfit. The definition of a nonfit will depend on the statistical framework. For the regression case, it has already been defined and for the location case $\boldsymbol{\beta}$ is considered a nonfit for a given data set $Z_n \subset \mathbb{R}^p$ if it lies outside the convex hull of Z_n.

Location depth can also be used to estimate location. This estimator is called deepest location estimator and is defined as follows:

$$DL(Z_n) = \text{argmax}_{\beta} HD(\boldsymbol{\beta}, Z_n)$$

If there are more than one observation with maximal depth, then the deepest location will be the mean of those observations. Note that in the univariate case this estimator is equivalent to the median, and in the multivariate case the deepest location can be seen as a multivariate median, the Tukey median.

Hubert et al. (2001) explore the analogies between location and regression depth, and between deepest location and regression estimators.

6. Conclusions

In this work a brief introduction to the concept of data depth is presented. Its importance in the statistics field and its usefulness in practical applications are shown. The data depth concept allows the generalization of the concept of order to the multivariate case, a goal long pursued by statisticians. This notion of order in multivariate data sets enlarges the field of applications of multivariate analysis because it allows the non-parametric data treatment. Another major advantage of this concept is that most measures of depth are robust, making it particularly suitable for the treatment of real life data, where outliers are frequently present.

Some of the more advanced concepts relative to depth are difficult to handle. Mainly if depth is relative to a distribution in high dimensions. This text includes only initial simple ideas that are considered fundamental to uncover the complexity of the subject and its interest in practice. It is hoped that computational and theoretical progress will be made to support more practical applications. The current interest in this area and the ongoing studies make believe that the expectations will be attained.

References

Barnett, V. (1976). The ordering of multivariate data. *Journal of the Royal Statistal Society A*, **139**, 319-354.

Brown, B. and Hettmansperger, T. (1989). The affine invariant bivariate version of the sign test. *Journal of the Royal Statistal Society B*, **51**, 117-125.

Christmann, A. (2002). Classification based on the SVM and on regression depth. In Dodge, Y. (editores), *Statistical data analysis based on the L1 norm and related methods*, Basel: Birkhöuser, 341-352.

Donoho, D. L. and Gasko, M. (1992). Breaking Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness, *Annals of Statistics*, **20**, 1803-1827.

Fraiman, R., Liu, R. and Meloche, J. (1997). Multivariate density estimation by probing depth. In *L1-Statistical Procedures and Related Topics*. Hayward, CA: IMS, 415-430.

Hubert, M., Rousseeuw, P. and Van Aelst, S. (2001). Similarities between location depth and regression depth. In Fernholz, L., Morgenthaler, S. and Stahel, W. (editores), *Statistics in Genetics and in the Environmental Sciences*. Basel: Birkhäuser, 159-172.

Hettmansperger, T., Nyblom, J. and Oja, H. (1992). On multivariate notions of sign and rank. In Dodge, Y. (editores), *L-1 Statistics and Related Methods*. Amsterdam: Elsevier , 267-278.

Hettmansperger, T. and Oja, H. (1994). Affine invariant multivariate multisample sign tests. *Journal of the Royal Statistal Society B*, **56**, 235-249.

Jörnsten, R. (2004). Clustering and classification based on the L1 data depth. *Journal of Multivariate Analysis*, **90**, 67-89.

Liu, R. (1990). On a notion of data depth based on random simplices. *Annals of Statistics*, **18**, 405-414.

Liu, R., Parelius, J. M. and Singh, K. (1999). Multivariate analysis by data depth descriptive statistics, graphics and inference. *Annals of Statistics*, **27**, 783-858.

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., and Hubert, M. (1999). Regression depth (with discussion). *Journal of the American Statistical Association*, **4**, 388-433.

Rousseeuw, P., Ruts, I. and Tukey, J. (1999). The Bagplot: A bivariate boxplot. *The American Statistician*, **53(4)**, 382-387.

Ruts, I. and Rousseeuw, P. (1996). Computing depth contours of bivariate point clouds. *Computational Statistics and 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.

Serfling, R. and Zuo, Y. (2000). General notions of statistical depth function. *Annals of Statistics*, **28**, 461-482.

Tukey, J. (1975). Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians*, **2**, Vancouver, 523-531.

Van Aelst, S., Van Driessen, K. and Rousseeuw, P. (2000). A robust method for multivariate regression. In Kiers, H. A. L. (editors), *Data analysis, classification, and related methods*. Berlin: Springer, 309-315.

Van Aelst,S. and Rousseeuw, P. (2000). Robustness of deepest regression. *Journal of Multivariate Analysis*, **73**, 82-106.

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.