Aalto-University
School of Science
Degree program of Engineering Physics and Mathematics

Local depth functions

MS-E2108 Independent Research Projects in Systems Analysis

Pekka Alli 20.9.2016

Instructor: Assistant professor Pauliina Ilmonen

Supervisor: Assistant professor Pauliina Ilmonen

Contents

1 Introduction

Statistical depth functions were introduced to extend the concept of median to multivariate data. In general, centrality of data is not a trivial measure, unlike in one dimensional case. Moreover, depth function provide usefull information about underlying distribution of data such as spread, shape and symmetry. One most vital property of depth function is its ability to provide center outward ordering of data, from the deepest points to the most shallow points. [1]

Depth functions are popular nonparametric way to describe data and therefore common concept in classification. Depth functions are used to determine which distribution generated a sample point using max-depth classification. In other words, measuring how central a sample point is respect to a given empiric distribution. In addition, depth function are used in outlier detection and testing normality assumptions of data. [2]

There is a fairly new concept of depth function, local depth. It is sometimes desirable to measure the depth in a certain neighbourhood, instead of global environment. This approach has some advantages, for example in classification where local values have sometimes dominating effect. Local depth function were introduced earliest at year 2011, but not in a generic way. The introduced depth function extensions were specific for each depth function and not inline with desired properties of statistical depth function. This study concentrates to the new study which defines local depth functions in a generic way to any locality level for any (global) depth function. [3]

Structure of this study is following: first we take a look at the definition of depth function. We display its requirements and common properties. After definitions we show some examples of common depth functions. Furthermore, we show how depth functions are used in classification. Afterwards we show required definitons for local depth functions, which are depth-region neighbourhoods. Then, we show the way to build depth function with arbitary locality level from any given depth function. Finally we show a simulation of use of a local depth function.

2 Depth functions

2.1 Definitions

Depth function were introduced to extend the concept of median to multivariate data and it measures centrality of given input x respect to distribution P. Statistical depth function is any function which provide P-based center-outward ordering of $x \in \mathbb{R}^d$ respect to the deepest points and fulfills following requirements: affine invariance, maximality at center, monotonicity relative to deepest point and vanishing at infinity. These requirements are defined below in (P1) - (P4). [1]

Let x be in $\mathbb{R}^d, d \in \mathbb{N}$ and P any statistical distribution. First requirement for statistical depth function can be formulated as following:

(P1) Statistical depth function $D(x, P)$ should be affine invariant. In other-words, the result should not be affected by choise of the coordinate system:

$$D(x, P) = D(Ax + b, P^{Ax+b}), \tag{1}$$

where A is nonsingular $d \times d$ matrix, and b is $d \times 1$ vector. P_{Ax+b} is the distribution for $Ax + b$ when x is having distribution P.

Second requirement for depth function is that it achieves its maximal value at the center.

(P2) If there is a unique point of symmetry θ, such that $\theta + x$ is distributed as $\theta - x$, then $D(\theta, P) = \sup D(x, P)$.

Third requirement is monotonicity of depth function respect to the depeest point.

(P3) If there exists deepest point θ, then for any vector v $D(\theta + t \cdot v)$ is monotonically decreasing function of $t > 0$.

Fourth requirement is that depth function vanishes at infinity.

(P4) Limit of depth function equals zero when the euclidian norm of the argument x approaches infinity:

$$\lim_{||x|| \to \infty} D(x, P) = 0, \tag{2}$$

where $|| \cdot ||$ denotes euclidian norm.

Properties (P1) - (P4) are properties that ideal depth function should possess but they are not necessary [1]. There are fairly popular depth functions which violate some of the previous requirements. Other desired properties of depth functions are that they are bounded and nonnegative. [1]

In real world, while speaking of distributions, the distribution usually consist of sample points instead of theoretical distribution. Therefore it is desired to define depth functions and it's requirements to empirical distributions.

Let us denote the empirical measures of distribution by $P^{(n)}$. The sample version of depth function $D(x, P)$ is denoted by $D(x, P^{(n)})$. Previous definitions (P1) - (P4) can be expaned to a empirical situation by replacing P with it's empirical version $P^{(n)}$.

2.2 Common depth functions

There are some statistical depth function which have gained popularity in use. To list few of most popular ones:

Half-space depth D_H is defined as following for distribution P:

$$D_H(x, P) = \inf_{u \in S^{d-1}} P(u^t(X - x) \geq 0), \tag{3}$$

where $X \sim P$ and $S^{d-1} = \{u \in \mathbb{R}^d : ||u|| = 1\}$. The half-space depth is smallest propability of half-spaces containing sample point x. This depth function fulfills properties (P1) - (P4). [4]

Simplicial depth D_S is defined as following

$$D_S(x, P) = P[x \in S(X_1, X_2, ..., X_{d+1})], \tag{4}$$

where $S(X_1, X_2, ..., X_d + 1)$ is closed simplex with verices $[x_1, x_2, x_3, ..., x_{d+1}]$ and $X_1, X_2, ...$ are i.i.d from P. This depth function does not fulfill all the proposed properties.

Mahalanobish depth D_M is defined as

$$D_M(x, P) = \frac{1}{1 + d^2_{\Sigma}(p)(x, \mu(P))}, \tag{5}$$

where μ and Σ are location and scatter functionals.

2.3 Depth function in classification

2.3.1 Classification

Classification is a task of predicting the class of a new obsevation x with respect to given distributions P_j. In other words, classification tries to identify to which distribution (category) a new sample point belongs. Classification of multivariate data is a common supervised learning problem and has applications in countless fields. An example would be classifying patients diseases (flu, zikavirus, pneumonia) by its symptons (blood pressure, body temperature, age). In this example the set of diseases (flu, zikavirus, pneumonia) are classes and symptoms (blood pressure, body temperature, age) are explanatory variables. Explanatory variables can be numerical, categorical or ordinal-valued. [2]

In classification we are mapping a new value x to a predictor for corresponding class Y. This mapping is done with a classifier. Classifier is an implementation of classification function $m : \mathbb{R} \rightarrow 0, 1$ for any value $\mathbf{x}$ and for any class Y. Classifiers can be categorized by their approaches, where some of the popular ones are: Linear classifiers, Support Vector Machines, Bayesian classifiers, Neural networks, Kernel Estimators and Depth-based classifiers.

Classification process usually contains steps i) training ii) validation and iii) prediction.

Training is only for parametric methods. In this process where the algorithm learns its parameters by training data.

Validation of the method contains calculation of misclassification rate of given algorithm. Misclassification rate is the ration of predicted wrong outputs on number of sample points. Different misclassifications can be given weight while calculating the ration. This is due that sometimes false negatives are greatly more harmful than false positives. For example while diagnosing deadly disease, it would be less harmfull to predict false positive than other way around. Misclassification given by validation with test set underestimates greatly the real misclassification rate. This can lead to false confidence even in cases with large test set.

Validation phase is usually intended to select the method for the classification task. As there are no unique best classifier for arbitary situation, classification should have a phase for selecting the applied method. There are different kind of approaches for validation. One highly used method is cross validation, which calculates misclassification rates for all training-test-set ratios chosen.

The last phase of classification is the prediction of the class of the new observation. This consist of appliying the chosen classifier to the new observation and interpreting the outcome.

2.3.2 Max-depth classification

Statistical depth function are in a key position for task of classifying. Max-depth classification is nonparametric in nature. In max-depth based classification, for given distributions P_i, $i \in 0, 1$ and a sample point x, the sample point x belongs most probably to the distributions which maximises chosen depth function $D(x, P_i)$.

Max-depth classifier can be presented in a form of

$$m_D(x) = \mathbb{I}[D(x, P_1) > D(x, P_0)] \tag{6}$$

where $\mathbb{I}$ denotes indicator function

$$\mathbb{I}(x) = \begin{cases} 1 & x \quad is \quad true \\ 0 & otherwise. \end{cases} \tag{7}$$

The distribution P_i can also be replaced with its empirical version and modify the classifying problem to answer to which population sample point x belongs to instead of from which distribution its from.

Although depth functions are great nonparametric way to define data, there is one drawback in its use as a classifier. Classifier should not require any special assumptions of the structure of the given data. Depth functions unfortunetly provide consistent results under for elliptical distribution only. Consistency in context of classification means that the classifier provides results that approach best possible classification rate as the test sets size increases.[3]

2.3.3 K-nearest neigbours

Classification with depth function can be done from two different aspects: based on depth function value (max-depth classification) and based on k-nearest neightbourhood. The neighbourhood is defined in sense of depths.

To define neighbourhood, the order of points have to defined around the sample point so that k nearest points can be found. In k-nearest neighbour classification, the class of sample is defined by the class that of the most of the neighbours have. Depth functions can be used to arrange the order of the points so that the k nearest neighbours for a new sample point can be defined in sense of depth.

3 Local depth functions

Depth functions usually analyse global centrality in the whole dataset. The introduced concept of local depth defines the concept of centrality in local environment instead of global. [5]

Some proposals of local depth functions fails to describe centrality when local measures converge to extereme. Therefore, a new approach is introduced to describe an extension of centrality at any level of locality still remaining to describe the centrality of the data. The new approach allows to turn any global depth function to corresponding local version of the function.

This approach uses neighbourhoods defined in next section.

3.1 Depth-based neighborhoods

Depth regions are defined for level α as a set that fulfills following property:

$$R_\alpha(P) = \{x \in \mathbb{R}^d | D(x, P) \geq \alpha\}. \tag{8}$$

For chosen α it is set of all points where depth function is greater than α. These regions are nested so that set with larger α is inside all sets with smaller α.

These sets called the depth region of order α can be indexed with probability parameter $\beta \in [0, 1]$ in such a way that

$$R^\beta(P) = \cap_{\alpha \in A(\beta)} R_\alpha(P), \tag{9}$$

such that

$$A(\beta) = \{\alpha \geq 0 : P(R_\alpha(P)) \geq \beta\}. \tag{10}$$

This means a region of smallest depth region of order α that has P-probability larger than or equal to β. Regions $R_\alpha(P)$ and $R^\beta(P)$ defines regions of around the deepest points only.

As depth functions provide center-outward ordering, with respect to the deepest points, this results that $X_i' s$ can be ordered in such fashion

$$D(X_1, P^{(n)}) \geq D(X_2, P^{(n)}) \geq ... \geq D(X_n, P^{(n)}), \tag{11}$$

where indices reflect the order of point's distantce from corresponding deepest points and $P^{(n)}$ is empirical distribution. As the order of $X_i' s$ can be measured with statistical depth function, this means that depth function can define a neighborhood around the deepest point $\theta^{(n)}$ in sense on k-nearest neighbors.

The previous definitions of depth based neighborhoods are respect to the deepest point only and beign global in that sense. To define local neightbourhood, we need depth based ordering respect to any point x.

Property (P2) of depth functions requires that center outward ordering should be symmetrical. To achieve local neighbouhood, the original distribution is symmetrized in respect to a local point x. This defines new distribution with original observation and their reflections respect to a local point x [3]

$$P_x = \frac{1}{2}P^X + \frac{1}{2}P^{2x-X}. \tag{12}$$

Other symmetrizations are proposed as well, as

$$P_x = \frac{1}{4}P^X + \frac{1}{4}P^{Rot_x^{\pi/2}(X)}\frac{1}{4}P^{Rot_x^{3\pi/2}(X)} + \frac{1}{4}P^{Rot_x^{\pi}(X)}, \tag{13}$$

where *Rot* stands for rotating variable x for given parameter. However, these other symmetrization fails in perspective of propertis (P1)-(P4) where this particular symmetrizatio yields non-affine invariant results.

This symmetrized version of P with regions $R^\beta(P_x)$ and $R_\alpha(P_x)$ are inline with most depth properties and define depth based neighbourhood around any point x.

This also leads to an ordering

$$D(X_{x,(1)}, P_x^{(n)}) \geq D(X_{x,(2)}, P_x^{(n)}) \geq ... \geq D(X_{x,(n)}, P_x^{(n)}) \tag{14}$$

which also defines k nearest neighbours of x.

3.2 From global depth to local depth

There are now two ways to define local neighborhood with depth functions, regions $R^\beta(P_x)$ and $R_\alpha(P_x)$. α-parametrization has some drawbacks. It is dependent on the distribution P and it is not always well defined. Therefore β-parametrization is used. It ranges from $\beta = 1$ no locality to $\beta = 0$ extreme locality. There were four ideal properties for depth functions. For local depths, these definitios are extended as following.

(P1') Affine-invariance, local depth function is invariant under coordinate change:

$$LD(Ax+b, P^{AX+b}, \beta) = LD(x, P, \beta), \tag{15}$$

(P2') Maximimality at any center of local symmetry. If θ is a β-local center of symmtery of F then

$$LD(\theta, F, \beta) = \sup_{x \in R^{\beta}(P_x)} LD(x, F, \beta) \tag{16}$$

(P3') at $\beta = 1$ monotonicity relative to the deppest point P3 is satisfied, (P4') vanishing at infinity:

$$\lim_{||x||->\infty} LD(x, F, \beta) = 0, \forall F, \forall \beta \in [0,1] \tag{17}$$

Any depth function can be transformed into a local version of a depth function via following formula:

$$LD^{\beta}(\cdot, P) : \mathbb{R}^d \to \mathbb{R}^+ : \mathbf{x} \to LD(\mathbf{x}, P) = D(\mathbf{x}, P_x^{\beta}), \tag{18}$$

where $P_x^{\beta} = P(\cdot | R_x^{\beta}(P))$ is conditional distribution of P, conditional on $R_x^{\beta}(P)$.

3.3 Sample local depth

In case of a sample $x_1, x_2, .., x_n$ i.i.d from common distribution P, we denote the empirical distribution $P^{(n)}$.

In global depth functions, the sample depth is aquired by changing the distribution P with the empirical distribution $P^{(n)}$ (with few modifications to the formulas).

With local depth this is also the case. We can define sample local depth

$$LD^{\beta}(\cdot, P^{(n)}) : \mathbb{R}^d \to \mathbb{R}^+ : \mathbf{x} \to LD(\mathbf{x}, P^{(n)}) = D(\mathbf{x}, P^{\beta_x,(n)}), \tag{19}$$

where $P^{\beta_x,(n)}$ are the points inside calculated region $R^{\beta}(P^{(n)})$.

As for sample local depth calculations, there is no best value of β-parameter. In practise it would be desirable to use few different values of β to see different results incase there is strong dependency between results and parameter β.

Sample local depth function can be proved to be consistent with corresponding theoretical depth functions as $n \to \infty$. Consistency theorem provides that for any absolutely continous P and any sequence $\beta_n -> \beta$

$$LD^{\beta_n}(x, P^{(n)}) \to LD^{\beta}(x, P) \quad as \quad n \to \infty. \tag{20}$$

This assumes that the given depth function $D(\cdot, P)$ is consistent for sample depth function for all absolutely continuous distribution P that converges weakly $P^{(n)} \to P$ when $n \to \infty$

$$\lim_{n->\infty} |D(x, P^{(n)}) - D(x, P)| = 0. \tag{21}$$

3.4 Earlier local depths

There have been ealier propositions of local depth functions. The previous local depth functions are not generic and have to defined separately for each depth function. Earlier local half-space depth function is defined as following [6]:

$$LHD(x0; F^X, \tau) = \inf_{u \in S_{p-1}} P(0 \leq u^T(X - x0) \leq \tau), \tag{22}$$

where τ defines the locality level.

Also for simplicial depth there is a previous definition of the local version. It maps the sample point to a propability of simplex with volume not greater than given locality parameter covering the sample point.

Both of the previous local depth functions are not affine-invariant and therefore won't fulfil the desired properties of depth function.

4 Simulation

4.1 Simulation of depth regions and symmetrization

In this chapter we simulate an example of use of depth based regions and symmetrization of distribution around a local sample point. In this simulation we have two populations from two different distributions P_1 and P_2.

Both population are samples from bivariate normal distributions with sample size of 2000. The depth function we use in this simulation is half-space depth.

Picture 1 shows us empirical distribution $P_1^{(n)}$ in red and empirical distribution $P_2^{(n)}$ in green. Single sample point is presented in blue. This picture shows that the sample points is around the center of the green points.

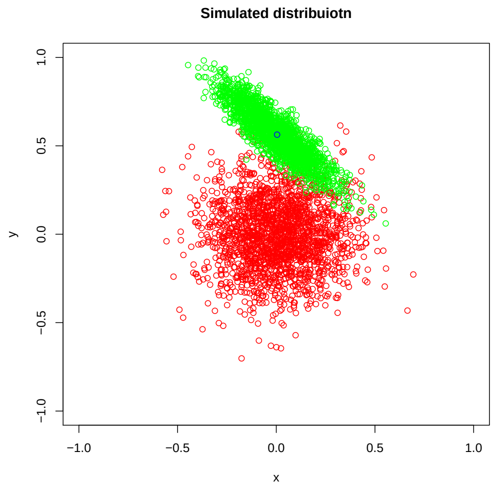

Figure 1: Empirical distributions. Distribution $P_1^{(n)}$ is in red and $P_2^{(n)}$ in green

First we demostrate the use of regions $R^\beta(P^{(n)})$ and $R_\alpha(P^{(n)})$. In picture 2 there are few different $R^\beta(P^{(n)})$ regions with different β values. $\beta = 0.4$ is colored in green, $\beta = 0.7$ is colored in yellow and $\beta = 1$ is red. Sample point is located in the $R^\beta(P^{(n)})|\beta = 0.4$ region of population P_2. So in this sense sample points are quite center respect to P_2.

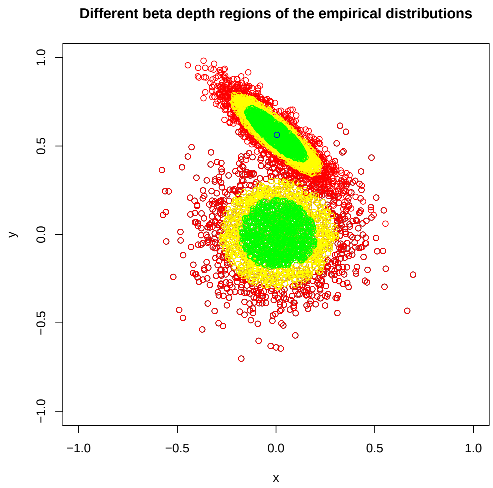

Figure 2: $R^{\beta}(P^{(n)})$ regions with different values of β

To compare these regions with $R_{\alpha}(P^{(n)})$ regions, in picture 3 there are regions $R_{\alpha}(P^{(n)})$ illustrated with corresponding values. The difference is that most of the data is left outside as the α always is dependent on the distribution and therefore the same values of β won't show any data necceserily. Almost all of the sample points are outside these regions of both populations.

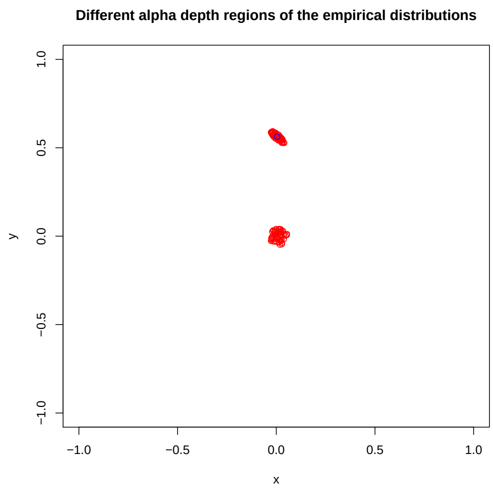

Figure 3: $R_{\alpha}(P^{(n)})$ regions with different values of α

Another demostration of $R_{\alpha}(P^{(n)})$ regions are illustrated in picture 4 with values of α 0.1, 0.2 and 0.4. This picture shows us more information than the previous.

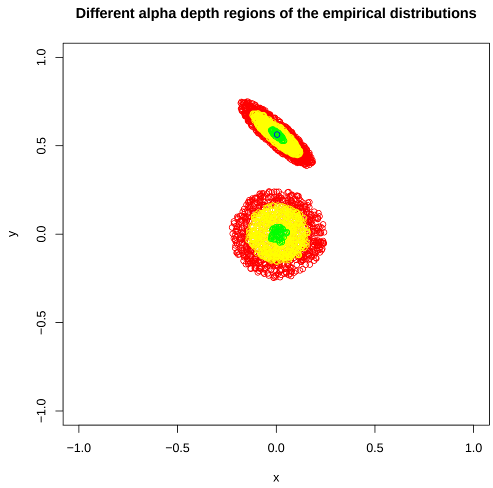

Figure 4: $R_\alpha(P^{(n)})$ regions with different values of α

To apply local depth function, the empirical distribution is symmerized as described in 12. To demostrate the symmetrization we show symmetrized distributions in respect to the sample point in picture 5. In the picture the sample point in blue.

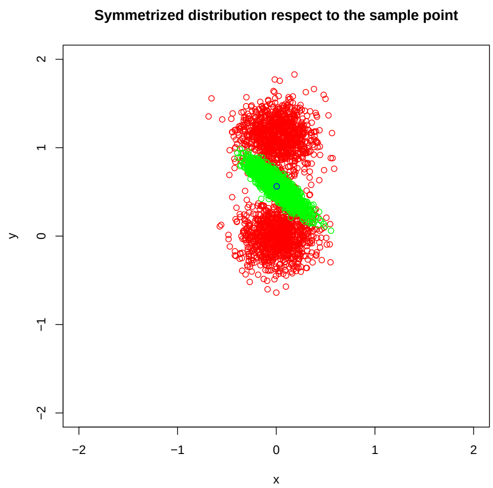

Figure 5: Distribution symmerized around the sample point

After symmetrization $R^\beta(P^{(n)})$ regions with $\beta = 0.9$ are displayed in picture 6.

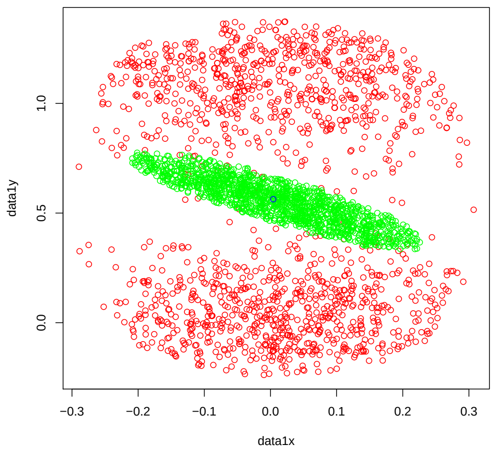

Figure 6: $R^\beta(P^{(n)})$ region of symmetric distribution with β value of 0.9

4.2 Simulation of local depth versus global depth

In this simulation we show the effect of chosen locality level. We have a empirical bivariate distribution $P = (P_x, P_y)$ which is simulated from distribution

$$P_x \sim \mathcal{U}(0,1) \tag{23}$$

$$P_y \sim \mathcal{U}(2(1-P_x^2), 5(1-P_x^2)), \tag{24}$$

where $\mathcal{U}$ denotes Uniform distribution. This simulation yields us moonshaped distribution which is shown in picture 7. In the picture we see a sample point in blue which is righ underneath the lower border of the distribution. The sample point is used to demostrate the effect of chosen locality level.

In picture 8 we have drawn $R^\beta(P^{(n)})$ with three different values of parameter β. Picture also contains the sample point. Region with $\beta = 1$ is in red, region with $\beta = 0.7$ is in yellow and region with $\beta = 0.4$ is in green. The regions are defined around the deepest point of the distribution. We can see from the picture that the sample point is near the green zone in sense of euclidian distance.

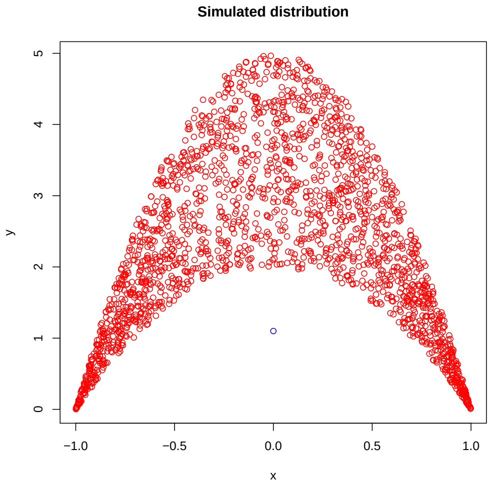

Figure 7: Simulated distribution

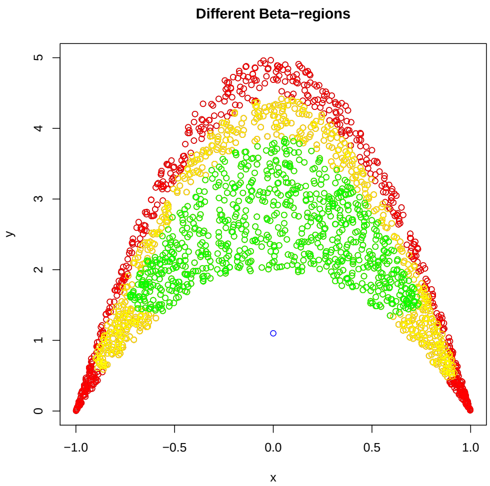

Figure 8: $R^{\beta}(P^{(n)})$ region of the empirical distribution

To see the effect of the locality level, have have measured the value of half-space depth function of the sample point in respect to the empirical distribution with different locality levels. The results are shown in picture 9. The results are that as the depth becomes more local, the centrality of the sample point respect to the distributions drops. This shows that in case of nonconvex distribution, the global depth fails to measure centrality properly.

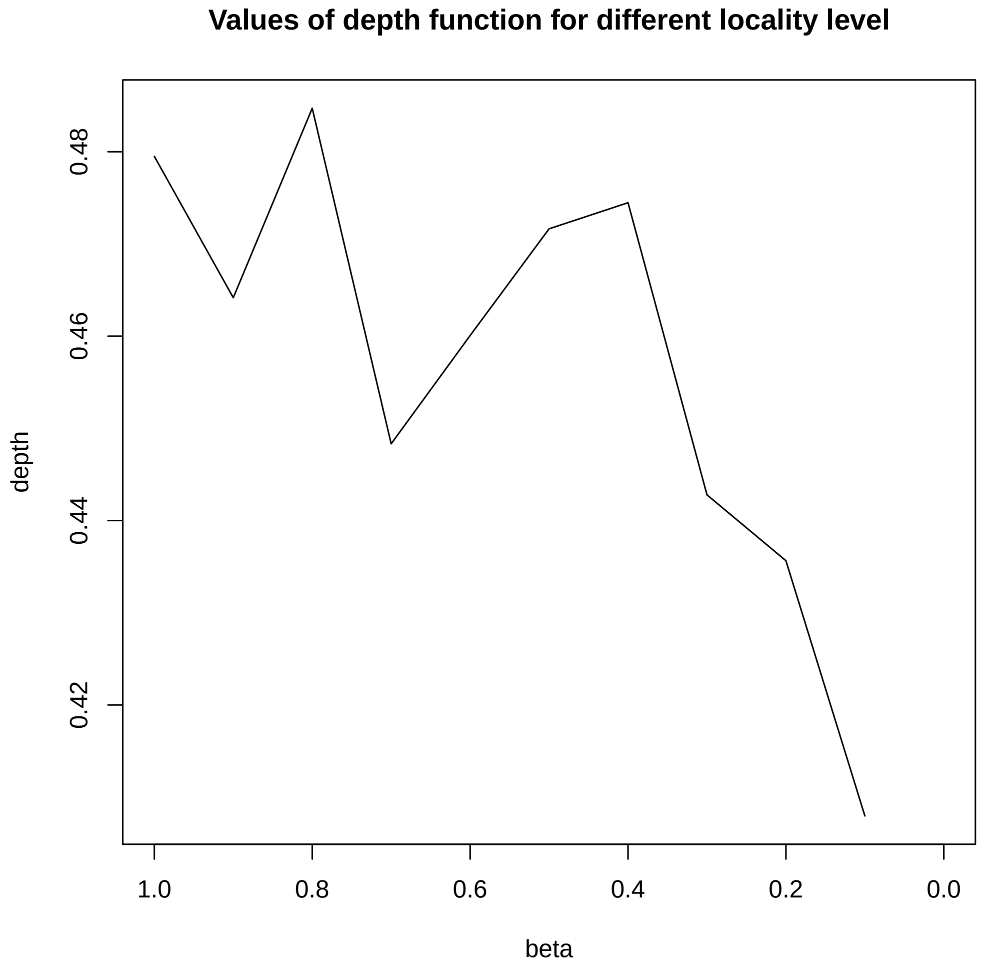

Figure 9: Half-space depth of the sample point respect to the empirical distribution

References

[1] Yijun Zuo and Robert Serfling. General notions of statistical depth function. *Ann. Statist.*, 28(2):461–482, 04 2000.

[2] Anil K Ghosh and Probal Chaudhuri. On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, 32(2):327–350, 2005.

[3] Davy Paindaveine and Germain Van Bever. Nonparametrically consistent depth-based classifiers. *Bernoulli*, 21(1):62–82, 2015.

[4] John W Tukey. Mathematics and the picturing of data. In *Proceedings of the international congress of mathematicians*, volume 2, pages 523–531, 1975.

[5] Davy Paindaveine and Germain Van bever. From depth to local depth: A focus on centrality. *Journal of the American Statistical Association*, 108(503):1105–1119, 2013.

[6] Claudio Agostinelli and Mario Romanazzi. Local depth. *Journal of Statistical Planning and Inference*, 141(2):817–830, 2011.