DepthProc: An R Package for Robust Exploration of Multidimensional Economic Phenomena

Daniel Kosiorowski
Cracow University of Economics

Zygmunt Zawadzki
Cracow University of Economics

Abstract

Data depth concept offers a variety of powerful and user friendly tools for robust exploration and inference for multivariate socio-economic phenomena. The offered techniques may be successfully used in cases of lack of our knowledge on parametric models generating data due to their nonparametric nature. This paper presents the R package **DepthProc**, which is available under GPL-2 licence on CRAN and R-forge servers for Windows, Linux and OS X platform. The package consist of among others successful implementations of several data depth techniques involving multivariate quantile-quantile plots, multivariate scatter estimators, multivariate Wilcoxon tests, robust regressions. In order to show the package capabilities, real dataset concerning *United Nations Fourth Millennium Goal* is used.

Keywords: Statistical Depth Function, Robust Data Analysis, Multivariate Methods.

1. Introduction

The modern Economics crucially depend on advances in applications of multivariate statistics. We mean here for example theory and practice of the portfolio optimisation, a practice of credit scoring, evaluation of results of government aid programs, creation of a taxation system or assessment of attractiveness of candidates on a labour market.

Unfortunately, in the Economics we very often cannot use powerful tools of the classical multivariate statistics basing on the mean vector, the covariance matrix and the normality assumptions. In a great part, the economic phenomena departure from normality. Usually our knowledge of the economic laws is not sufficient for a parametric modelling. Moreover, the today Economics significantly differs from a tomorrow Economics due to technological development and/or an appearance of new social phenomena. Additionally the data sets under our consideration consist of outliers and or inliers of various kind and/or we have to cope with a missing data phenomenon.

Robust statistics aims at identifying a tendency represented by an influential majority of data and detecting observations departing from that tendency (see Maronna, Martin, and Yohai (2006)). Nonparametric and robust statistical procedures are especially useful in the Economics where an activity of influential majority of agents determines behaviour of a market, closeness to a crash etc. From a conceptual point of view, robust statistics is closely tied with well known economic ideas like *Pareto's effectiveness* or *Nash equilibrium* (see Mizera (2002)).

The main aim of this paper is to present an R package (R Core Team (2013)) **DepthProc** consisting of successful implementations of a selection of multivariate nonparametric and robust procedures belonging to so called *Data Depth Concept* (DDC), which are especially useful in exploration of socio-economic phenomena. The package is available under GPL-2 license on **CRAN** and **R-forge servers**.
The rest of the paper is organized as follows: in Section 2, basic notions related to the data depth concept are briefly described. In Section 3, the procedures offered by the package are briefly presented. In Section 4, an illustrative example is presented. The paper ends with some conclusions and references. All data sets and examples considered within the paper are available after installing the package.

In this paper we use the following notation and definitions borrowed from Dyckerhoff (2004). S^{d-1} is the $(d-1)$ dimensional unit sphere in $\mathbb{R}^d$, $S^{d-1} = \left\{x \in \mathbb{R}^d : \|x\| = 1\right\}$. $\mathcal{B}^d$ denotes Borel σ algebra in $\mathbb{R}^d$. The transpose of a vector $x \in \mathbb{R}^d$ is written by $x^\top$. For a random variable X we write Q_X for the usual (lower) quantile function, $Q_X : (0,1) \to \mathbb{R}$, $Q_X(p) = \min\{x \in \mathbb{R}^d : P(X \leq x) \geq p\}$, and $\bar{Q}_X$ for the upper quantile function $\bar{Q}_X : (0,1) \to \mathbb{R}$, $\bar{Q}_X(p) = \max\{x \in \mathbb{R}^d : P(X \geq x) \geq p\}$. A sample consisting of n observations is denoted by $X^n = \{x_1, ..., x_n\}$, F denotes a probability distribution in $\mathbb{R}^d$, and F_n its empirical counterpart.

2. Data depth concept

Data depth concept was originally introduced as a way to generalize the concepts of median and quantiles to the multivariate framework. A detailed presentation of the concept can be found in Liu, Parelius, and Singh (1999), Zuo and Serfling (2000), Serfling (2003), Serfling and Wang (2006), and Mosler (2013). Nowadays the DDC offers a variety of powerful techniques for exploration and inference on economic phenomena involving robust clustering and classification, robust quality control and streaming data analysis, robust multivariate location, scale, symmetry tests. Theoretical aspects of the concept could be found for example in Kong and Zuo (2010) and in references therein, recent developments of the computational aspects presents for example Shao and Zuo (2012). Our package **DepthProc** uses so called location depths and their derivatives, i.e., regression depth and Student depth. The **DepthProc** implements also recently developed concept of local depth presented in Paindavaine and Van Bever (2012) and Paindavaine and Van Bever (2013). A developer version of the package, which is available on **R-forge** servers, additionally consists of fast algorithms for calculating selected depths for functional data and weighted by the local depth nonparametric estimators of a predictive distribution.

2.1. Basic definitions

Following Dyckerhoff (2004) we consider the depth of a point w.r.t. a probability distribution. Let $\mathcal{P}_0$ be the set of all probability measures on $(\mathbb{R}^d, \mathcal{B}^d)$ and $\mathcal{P}$ a subset $\mathcal{P}_0$. A depth assigns to each probability measure $F \in \mathcal{P}$ a real function $D(\cdot, F) : \mathbb{R}^d \to \mathbb{R}_+$, the so-called depth function w.r.t. F.

The set of all points that have depth at least α is called $\alpha-$ **trimmed region**. The $\alpha-$

trimmed region w.r.t. F is denoted by $D_\alpha(F)$, i.e.,

$$D_\alpha(F) = \left\{ z \in \mathbb{R}^d : D(z, F) \geq \alpha \right\}. \tag{1}$$

In a context of applications, the probability measure is the distribution F^X of a d-variate random vector X. In this case we write shortly $D(z, X)$ instead of $D(z, F^X)$ and $D_\alpha(X)$ instead of $D_\alpha(F^X)$. The data depth is then defined on the set $\mathcal{X}$ of all random vectors X for which F^X is in $\mathcal{P}$.
Formal definitions of the depth functions can be found in Liu *et al.* (1999), Zuo and Serfling (2000), Mosler (2013). There is an agreement in the literature, that every concept of depth should satisfy some reasonable properties:

T1 *Affine invariance*: For every regular $d \times d$ matrix A and $b \in \mathbb{R}^d$ it holds $D(z, X) = D(az + b, AX + b)$.

T2 *Vanishing at infinity*: For each sequence $\{x_n\}_{n\in\mathbb{N}}$ with $\lim_{n\to\infty} \|x_n\| = \infty$ holds $\lim_{n\to\infty} D(x_n, X) = 0$.

T3 *Upper semicontinuity*: For each $\alpha > 0$ the set $D_\alpha(X)$ is closed.

T4 *Monotone on rays*: For each x_0 of maximal depth and each $r \in S^{d-1}$, the function $\mathbb{R}_+ \to \mathbb{R}$, $\lambda \mapsto D(x_0 + \lambda r, X)$ is monotone decreasing.

$T4^*$ *Quasiconcavity*: For every $\alpha \geq 0$ holds: If z_1, z_2 are two points with a depth of at least α , then every point on the line segment joining z_1 and z_2 has depth of at least α , too.

DEFINITION (Dyckerhoff (2004)): A mapping D , that assigns to each random vector X in a certain set $\mathcal{X}$ of random vectors a function $D(\cdot, F) : \mathbb{R}^d \to \mathbb{R}_+$ and that satisfies the properties T1, T2, T3 and T4 is called depth. A depth that satisfies $T4^*$ is called convex depth

Properties T1 to T4 are formulated in terms of the depth itself. It is very useful to notice however, that these properties can also be formulated in terms of the trimmed regions (what is useful for approximate depth calculation):

Z1: *Affine equivariance*: For every regular $d \times d$ matrix A and $b \in \mathbb{R}^d$ it holds $D_\alpha(AX + b) = AD_\alpha(X) + b$.

Z2: *Boundedness*: For every $\alpha > 0$ the $\alpha-$ trimmed region $D_\alpha(X)$ is bounded.

Z3: *Closedness*: For every $\alpha > 0$ the $\alpha-$ trimmed region $D_\alpha(X)$ is closed.

Z4: *Starshapedness*: If x_0 is contained in all nonempty trimmed regions, then the trimmed regions $D_\alpha(X)$, $\alpha \geq 0$, are starshaped w.r.t. x_0 .

$Z4^*$: *Convexity*: For every $\alpha > 0$ the $\alpha-$ trimmed region $D_\alpha(X)$is convex.

Z5: *Intersection property*: For every $\alpha > 0$ holds $D_\alpha(X) = \bigcap_{\beta:\beta<\alpha} D_\beta(X)$.

The simplest example of the depth is **the Euclidean depth** defined as (see 11)

$$D_{EUK}(y, X^n) = \frac{1}{1 + \|y - \bar{x}\|^2}, \tag{2}$$

where $\bar{x}$ denotes the mean vector calculated from a sample X^n.

As a next example let us take **the Mahalanobis depth** (see Fig. 3)

$$D_{MAH}(y, X^n) = \frac{1}{1 + (y - \bar{x})^\top S^{-1}(y - \bar{x})}, \tag{3}$$

where S denotes the sample covariance matrix X^n.

A **symmetric projection depth** $D(x, X)$ of a point $x \in \mathbb{R}^d$, $d \geq 1$ is defined as

$$D(x, X)_{PRO} = \left[1 + sup_{\|u\|=1} \frac{\left|u^\top x - Med\left(u^\top X\right)\right|}{MAD\left(u^\top X\right)}\right]^{-1}, \tag{4}$$

where Med denotes the univariate median, $MAD(Z) = Med(|Z - Med(Z)|)$. Its sample version denoted by $D(x, X^n)$ or $D(x, X^n)$ is obtained by replacing F by its empirical counterpart F_n calculated from the sample X^n (see Fig. 1). This depth is affine invariant and $D(x, F_n)$ converges uniformly and strongly to $D(x, F)$. The affine invariance ensures that our proposed inference methods are coordinate-free, and the convergence of $D(x, X^n)$ to $D(x, X)$ allows us to approximate $D(x, F)$ by $D(x, X^n)$ when F is unknown. Induced by this depth, multivariate location and scatter estimators have high breakdown points and bounded Hampel's influence function (for further details see Zuo (2003)).

Next, very important depth is **the weighted L^p depth**. The weighted L^p depth $D(\mathbf{x}, F)$ of a point $\mathbf{x} \in \mathbb{R}^d$, $d \geq 1$ generated by d dimensional random vector $\mathbf{X}$ with distribution F, is defined as (see Fig. 4)

$$D(x, F) = \frac{1}{1 + Ew(\|x - X\|_p)}, \tag{5}$$

where w is a suitable weight function on $[0, \infty)$, and $\|\cdot\|_p$ stands for the L^p norm (when $p = 2$ we have usual Euclidean norm). We assume that w is non-decreasing and continuous on $[0, \infty)$ with $w(\infty-) = \infty$, and for $a, b \in \mathbb{R}^d$ satisfying $w(\|a + b\|) \leq w(\|a\|) + w(\|b\|)$. Examples of the weight functions are: $w(x) = a + bx$, $a, b > 0$ or $w(x) = x^\alpha$. The empirical version of the weighted L^p depth is obtained by replacing distribution F of X in $Ew(\|x - X\|_p) = \int w(\|x - t\|_p) dF(t)$ by its empirical counterpart. The weighted L^p depth from sample $X^n = \{x_1, ..., x_n\}$ is computed as follows:

$$D(x, X^n) = \frac{1}{1 + \frac{1}{n}\sum_{i=1}^{n} w\left(\|x - X_i\|_p\right)}, \tag{6}$$

The weighted L^p depth function in a point, has the low breakdown point (BP) and unbounded influence function IF (see Maronna *et al.* (2006) for the BP and IF definitions). On the other hand, the weighted L^p depth induced medians (multivariate location estimator) are globally robust with the highest BP for any reasonable estimator. The weighted L^p medians are also locally robust with bounded influence functions for suitable weight functions. Unlike other

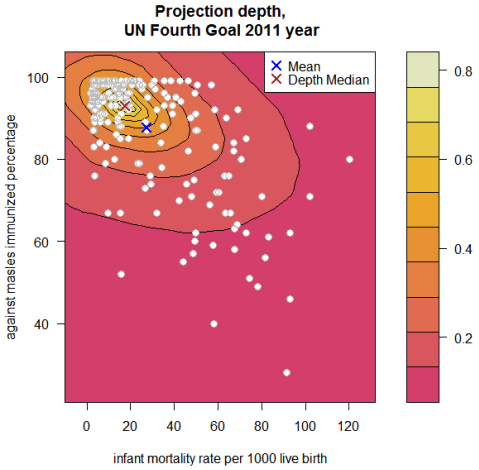

Figure 1: Projection depth.

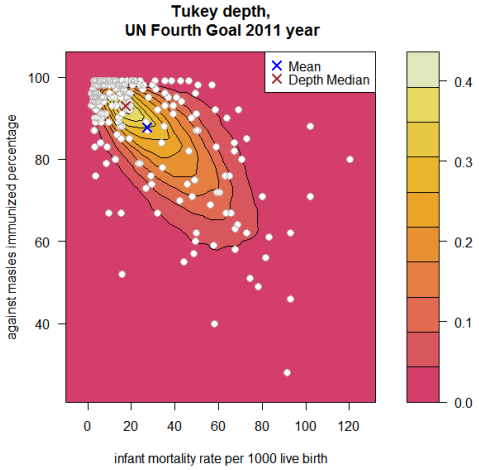

Figure 2: Tukey depth.

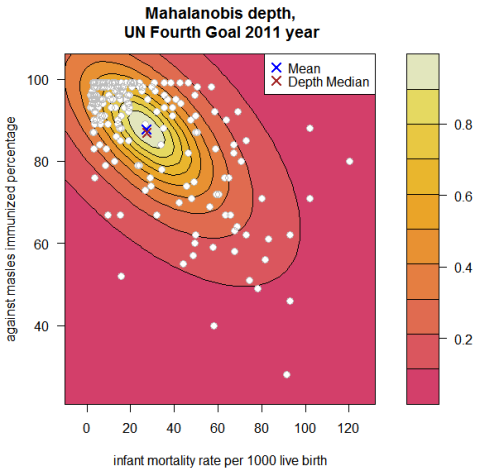

Figure 3: Mahalanobis depth.

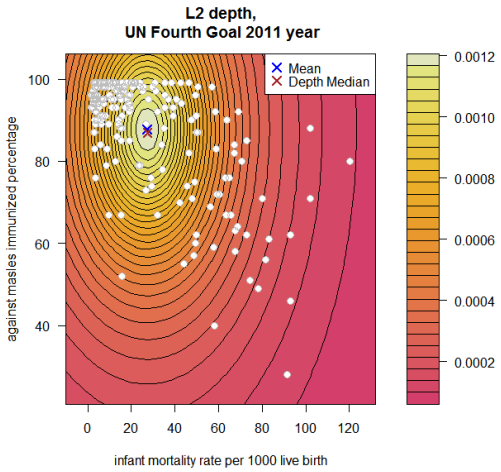

Figure 4: L^2 depth.

existing depth functions and multivariate medians, the weighted L^p depth and medians are computationally feasible for on-line applications and easy to calculate in high dimensions. The price for this advantage is the lack of affine invariance and equivariance of the weighted L^p depth and medians, respectively. Theoretical properties of this depth can be found in Zuo (2004).

Next, very important depth is **the halfspace depth** (Tukey depth, see Fig. 2)

$$D(x,F) = \inf_{H} \left\{ P(H) : x \in H \subset \mathbb{R}^d, \text{ X is closed subspace} \right\} \tag{7}$$

A very useful for the economic applications example of depth, originating from the halfspace depth, is **regression depth** introduced in Rousseeuw and Hubert (2004) and intensively studied in Van Aelst and Rousseeuw (2000) and in Mizera (2002).

Let $Z^n = \{(x_1, y_1), ..., (x_n, y_n)\} \subset \mathbb{R}^d$ denotes a sample considered from a following semipara-

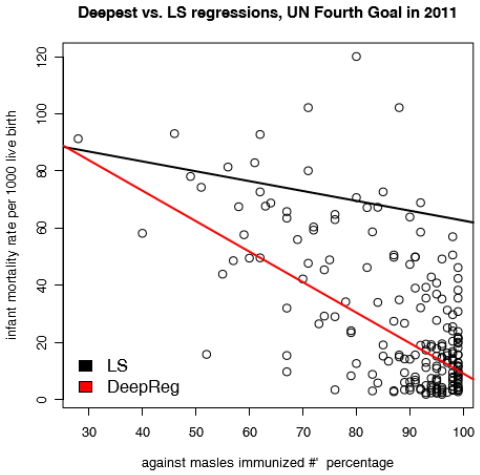

Figure 5: Deepest regression.

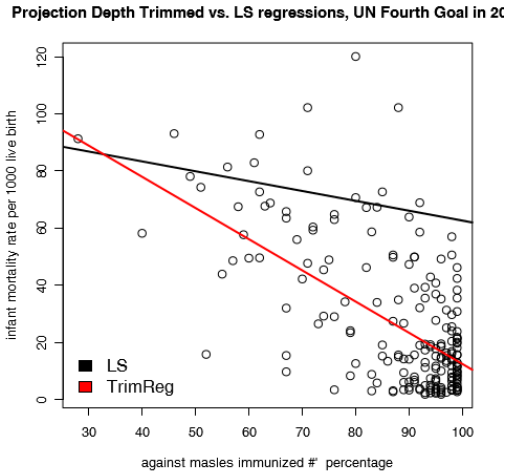

Figure 6: Depth trimmed regression.

metric model:

$$y_l = a_0 + a_1 x_{1l} + ... + a_{(d-1)l} x_{(d-1)l} + \varepsilon_l, l = 1, ..., n, \tag{8}$$

we calculate a depth of a fit $\alpha = (a_0, ..., a_{d-1})$ as

$$RD(\alpha, Z^n) = \min_{u \neq 0} \sharp \left\{ l : \frac{r_l(\alpha)}{u^\top x_l} < 0, l = 1, ..., n \right\}, \tag{9}$$

where $r(\cdot)$ denotes the regression residual, $\alpha = (a_0, ..., a_{d-1})$, $u^\top x_l \neq 0$.

The deepest regression estimator $DR(\alpha, Z^n)$ is defined as

$$DR(\alpha, Z^n) = \arg\max_{\alpha \neq 0} RD(\alpha, Z^n) \tag{10}$$

Fig. 5 presents a comparison of least squares and DR estimators of simple regression. The regression depth has its local version thanks to its relation to the halfspace depth (see Paindaveine and Van Bever (2013)). The local version of this depth may be easily calculated within **DepthProc** package. Next depth, which is implemented within the package, is **the Student depth** originating from Mizera (2002) and which was proposed in Mizera and Müller (2004). It is pointed out in Mizera (2002) that general halfspace depth can be defined as a measure of data-analytic admissibility of a fit. Depth of the fit θ is defined as proportion of the observations whose omission causes θ to become *a nonfit*, a fit that can be uniformly dominated by another one.

For a sample $X^n = \{x_1, ..., x_n\}$ we consider a criterial function F_i, given a fit represented by α , the criterial function evaluates the lack of fit of α to the particular observation x_i . It means α^* fitting x_i better than α, if $F_i(\alpha^*) < F_i(\alpha)$.

In Mizera (2002) more operational version – the tangent depth of a fit α is defined

$$d(\alpha) = \inf_u \neq \mathbf{0} \left\{ \sharp n : u^\top \nabla_\alpha F_i(\alpha) \geq 0 \right\}, \tag{11}$$

where $\sharp$ stands for the relative proportion in the index set - its cardinality divided by n .

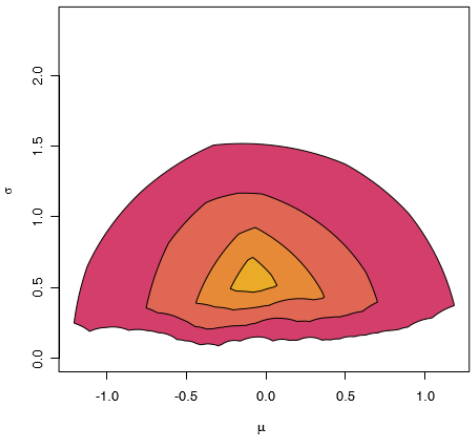

Figure 7: Sample student depth contour plot, data from N(0,1).

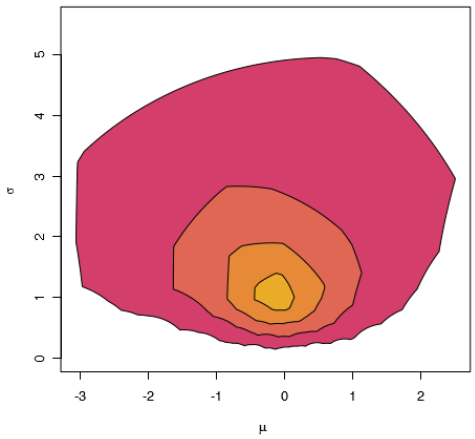

Figure 8: Sample student depth contour plot, data from t(1).

In Mizera and Müller (2004) authors suggest assuming the location-scale model for the data and taking log-likelihood in a role of the criterial function. They suggest taking the criterial function

$$F_i(\mu, \sigma) = -\log f\left(\frac{y_i - \mu}{\sigma}\right) + \log \sigma \tag{12}$$

Substituting into (12) into (11) we obtain a family of location-scale depths.

The Student depth of $(\mu, \sigma) \in \mathbb{R} \times [0, \infty)$ is obtained substituting into the above expression the density of the t distribution with v degrees of freedom

$$d(\mu, \sigma) = \inf_{u \neq 0} \left\{ \sharp i : (u_1, u_2) \begin{pmatrix} \tau_i \\ \frac{v}{v+1}(\tau_i^2 - 1) \end{pmatrix} \geq 0 \right\}, \tag{13}$$

where by the multiplication we mean the dot product,τ_i is a shorthand for $(y_i - \mu)/\sigma$, and we can absorb the constant $v/(v+1)$ into the u term (see Fig. 7 - 8)

The Student Median (SM) is the maximum depth estimator induced by the Student depth. It is very interesting joint estimator of location and scale in a context of robust time series analysis. It is robust but not very robust – its BP is about 33% and hence is robust to a moderate fraction of outliers but is sensitive to a regime change of a time series at the same time. It is worth noticing, that by its definition, the SM is not affected by temporal dependence of the observations.

2.2. Local depth

In an opposition to the density function, the depth function has a global nature i.e., e.g., that it expresses a centrality of a point w.r.t. a whole sample. This property is an advantage of depth for some applications but may be treated as its disadvantage in the context od classification of objects or for k-nearest neighbour rule applications. Depth based classifier or depth based k-nearest density estimators need local version of depths. A successful concept of **local depth** was proposed in Paindavaine and Van Bever (2012). For defining **a neighbourhood** of a

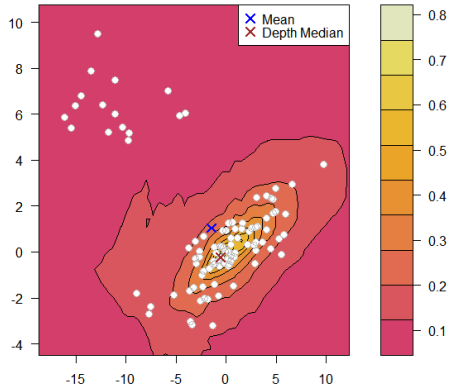

Figure 9: Local L^2 depth, locality = 60%.

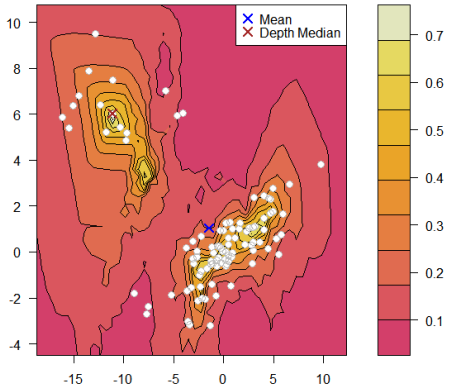

Figure 10: Local L^2 depth, locality = 20%.

point, authors proposed using idea of **symmetrisation** of a distribution (a sample) with respect to a point in which depth is calculated. In their approach instead of a distribution P^X, a distribution $P_x = 1/2P^X + 1/2P^{2x-X}$ is used.

For any $\beta \in (0, 1]$, let us introduce the smallest depth region with probability bigger or equal to β,

$$R^\beta(F) = \bigcap_{\alpha \in A(\beta)} D_\alpha(F), \tag{14}$$

where $A(\beta) = \{\alpha \geq 0 : P\left[D_\alpha(F)\right] \geq \beta\}$. Then for a locality parameter $\beta \in (0, 1]$ we can take a neighbourhood of a point x as $R^\beta(P_x)$ (see Fig. 9 - 10).

Formally, let $D(\cdot, P)$ be a depth function. Then the **local depth** with the locality parameter $\beta \in (0, 1]$ and w.r.t. a point x is defined as

$$LD^\beta(z, P) : z \to D(z, P_x^\beta), \tag{15}$$

where $P_x^\beta(\cdot) = P\left(\cdot | R_x^\beta(P)\right)$ is cond. distr. of P conditioned on $R_x^\beta(P)$.

For $\beta = 1$ the local depth reduces to its global counterpart (no localization). In a sample case $X^n = \{x_1, ..., x_n\}$, in a first step we calculate depth of a point y by adding to the original observations $x_1, ..., x_n$ their reflections $2y - x_1, ..., 2y - x_n$ w.r.t. y – let us denote this combined sample X_n^y and then calculating usual depth. Then we order observations from the original sample w.r.t. $D(\cdot, X_n^y)$ the sample depth calculated from the combined sample: $D(x_{(1)}, X_n^y) \geq ... \geq D(x_{(n)}, X_n^y)$. We choose the locality parameter $\beta \in (0, 1]$ determining a size of depth based neighbourhood of the point x. Then we determine $n_\beta(X_n^y) = \max\left\{l = \lceil n\beta \rceil, ..., n\right\} : D(x_{(l)}, X_n^y) = D(x_{(\lceil n\beta \rceil)}, X_n^y)\}$. Finally we calculate $LD^\beta(y, X^n) = D(y, X_n^{y,\beta})$, where $X_n^{y,\beta}$ denotes subsample $x_{(1)}, ..., x_{(n\beta)}$ of X_n^y. Further theoretical properties involving its weak continuity and almost sure consistency can be found in Paindavaine and Van Bever (2012) and Paindavaine and Van Bever (2013).

2.3. Approximate depth calculation

A direct calculation of many statistical depth functions is a very challenging computational issue. On the other hand a computational tractability of depths and induced by them procedures is especially important for online economy involving monitoring high frequency financial data, social networks or for shopping center management (see Kosiorowski (2014)).

Within the **DepthProc** package we use approximate algorithm proposed in Dyckerhoff (2004) to calculation of a certain class of location depth functions (depths possessing so called strong projection property), we base also on a algorithm proposed by Rousseeuw i Hubert (1998) for deepest regression calculation and direct algorithm **lsdepth** for Student depth calculation proposed in Müller (2003). For calculation of the local depths we use direct method described in Paindavaine and Van Bever (2012). Below we briefly present main ideas of the Dyckerhoff algorithm.

DEFINITION (Dyckerhoff (2004)): Let D be a depth on $\mathcal{X}$. D satisfies the (weak) projection property, if for each point $y \in \mathbb{R}^d$ and each random vector $X \in \mathcal{X}$ it holds:

$$D(y, X) = \inf \left\{ D(p^\top y, p^\top X) : p \in S^{d-1} \right\}.$$

THEOREM 1 (Dyckerhoff (2004)): For each $X \in \mathcal{X}$ let $(Z_\alpha(X))_{\alpha \geq 0}$ be a family of subsets of $\mathbb{R}^d$ that satisfy the properties Z1 to Z5. Further let $Z_0(X) = \mathbb{R}^d$ for every $X \in \mathcal{X}$. If D is defined by $D(z, X) = \sup \{\alpha : Z_\alpha(X)\}$, then D is a depth on $\mathcal{X}$ and the sets $Z_\alpha(X)$ are trimmed regions of D.

THEOREM 2 (Dyckerhoff (2004)): Let D^1 be a univariate depth. If D is defined by $D(z, X) = \inf_{p \in S^{d-1}} D^1(p^\top z, p^\top X)$, then D is a multivariate convex depth that satisfies the weak projection property.

Theorem 2 shows how multivariate depths can be obtained from univariate depths via the projection property. In theorem 1 a depth was defined by the family of its trimmed regions. By combining these two results one arrives at a construction method of multivariate depths from univariate trimmed regions. For practical applications of the above approach it is of prior importance to replace *sup* and *inf* by means of *max* and *min*, i.e., approximate multivariate depth by means of a finite number of projections. Theoretical background of the issue can be found in Cuesta-Albertos and Nito-Reyes (2008) and references therein.

In the **DepthProc** in order to decrease the computational burden related to sample depth calculation we use proposition 11 from Dyckerhoff (2004). We use 1000 random projections from the uniform distribution on a sphere. We use following families of one-dimensional central regions:

1. *For Tukey depth*

$$Z_\alpha(X) = \left[Q_X(\alpha), \bar{Q}_X(\alpha) \right],$$

2. *For zonoid depth (see also Mosler (2013), Lange, Mosler, and Mozharovskyi (2014))*

$$Z_\alpha(X) = \left[\tfrac{1}{\alpha} \int_0^\alpha Q_X(p)dp, \tfrac{1}{\alpha} \int_0^\alpha \bar{Q}_X(p)dp, \right],$$

3. *For a symmetric projection depth (see Zuo (2003))*

$$D_\alpha(X) = [med_X - c(\alpha)MAD_X, med_X + c(\alpha)MAD_X]\,, \text{ where } c(\alpha) = (1-\alpha)/\alpha.$$

2.4. Existing software for depth calculation

Currently there are four packages **CRAN** available which are directly dedicated for depth calculation: **depth** Genest, Masse, and Plante (2012), **depthTools** Lopez-Pintado and Torrente (2013), **localdepth** Agostinelli and Romanazzi (2013) and **ddalpha** Lange *et al.* (2014). Additionally, two packages **fda.usc** Febrero-Bande and de la Fuente (2012) and **fda** Ramsay, Hooker, and Graves (2009), consist of tools related to depths for functional data.

The **depth** package allows for exact and approximate calculation of Tukey, Liu and Oja depths. It also provides tools for visualisation contour plots and perspective plots of depth functions, and function for depth median calculation. Note, that commands `depthContour` and `depthPersp` which are available within the **DepthProc** were patterned on these **depth** commands.

The **depthTools** is focused on the Modified Band Depth (MBD) for functional data Lopez-Pintado and Romo (2009). It provides scale curve, rank test based on MBD and two methods of supervised classification techniques, the DS and TAD methods.

The **localdepth** package enables us for calculation of local version of "simplicial", "ellipsoid", "halfspace" (Tukey's depth), "mahalanobis" and "hyperspheresimplicial" depth functions. The **localdepth** also has a function for depth-vs-depth plot, which differs from the function which is available within the **DepthProc**. In the **localdepth**, the DDPlot is a plot of normalized localdepth versus normalized depth.We should note also that version of the local depth which is available within the **localdepth** differs from a more general version proposed in Paindavaine and Van Bever (2013), which is available within the **DepthProc**.

The **ddalpha** package concentrates around a new method for classification basing on the DD-plot prepared using the random Tukey depth and zonoid depth.

3. Package description and illustrative examples

Our package comprises among other of the commands listed in a Table 1. The `depthDensity` and `depthMBD` commands, dedicated correspondingly to nonparametric weighted by local depth conditional probability density estimator and for fast calculation of the modified band depth for functional data, are under development. These commands indicate a direction of a further development of the package however.

3.1. Available depths functions

A basic command for depth calculation is

```
depth(u, X, method = c("Projection", "Tukey","Mahalanobis","Euclidean","LP",
"local"), p=2, beta=0.5,...)
```

Arguments

u: Numerical vector or matrix, which depth is to be calculated. A dimension has to be the same as that of the observations.

Command	Short description
asymmetryCurve	multivariate asymmetry functional
binningDepth2d	depth based simple binning of 2D data
CovLP	L^p depth weighted location and scatter estimator
ddmvnorm	multivariate quantile-quantile normality plot
deepReg2d	deepest regression estimator for simple regression
depth	depth calculation
depthContour	depth contour plot
depthDensity	depth weighted density estimator
depthMBD	fast modified band depth calculation
depthmedian	multivariate median calculation
depthPersp	depth perspective plot
depthLocal	local depth calculation
lsdSampleMaxDepth	Student median calculation
medianDepthConfinterval	bootstrap region for a multivariate median
mWilcoxonTest	multivariate Wilcoxon test for location and/or scale differences
ScaleCurve	multivariate scatter functional
trimmReg2d	projection depth trimmed regression 2D

Table 1: Main commands available within the **DepthProc**.

X: The data as a matrix, a data frame or a list. If it is a matrix or data frame, then each row is treated as one multivariate observation. If it is a list, all components must be numerical vectors of equal length (coordinates of the observations).
method: Character string determining the depth function. The method can be "Projection" (the default), "Mahalanobis", "Euclidean", "Tukey", "LP" or "Local".
p: L^p depth parameter.
beta: locality parameter.

3.2. Maximal depth estimators

The **DepthProc** enables for calculating multivariate medians induced by depth functions.

```
depthMedian(x, ...)
```

Arguments:

x: The data as a $k \times 2$ matrix or data frame.
method: Character string determining the depth function. The method can be "Projection" (the default), "Mahalanobis", "Euclidean", "Tukey", "LP" or "Local".
p: L^p depth parameter.

3.3. DepthContour

Basic statistical plots offered by **DepthProc** are **the contour plot** and **the perspective plot** (see Fig. 11 – 12).

```
depthContour(x, n = 50, pmean = TRUE, mcol = "blue", pdmedian = TRUE,
```

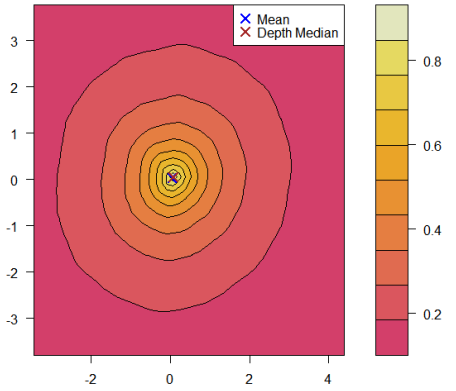

Figure 11: Sample contour plot.

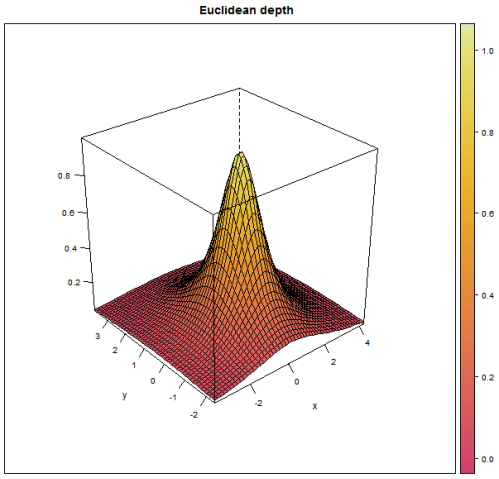

Figure 12: Sample perspective plot.

```
mecol = "brown", legend = TRUE, points = TRUE, xlab=" ", ylab=" ", main=" ",
method = c("Projection", "Tukey","Mahalanobis","Euclidean","LP","local"), p=2,
beta=0.5 )
```

Arguments

x: The data as a $k \times 2$ matrix or data frame.

3d Plot – by default, plot from lattice is drawn. You can use plot_method="rgl", but currently **rgl** is not on "depends" - list. Note - **rgl** can cause some problems with installation on clusters without OpenGL.

```
depthPersp(x, plot_method = "lattice", xlim = extendrange(x[, 1], f = 0.1),
ylim = extendrange(x[, 2], f = 0.1), n = 50, xlab = "x", ylab = "y", plot_title
= NULL, ...)
```

Arguments

x: The data as a $k \times 2$ matrix or data frame.

3.4. DD-plots

For two probability distributions F and G, both in $\mathbb{R}^d$, we can define **depth vs. depth** plot being very useful generalization of the one dimensional quantile-quantile plot:

$$DD(F,G) = \left\{ \left(D(z,F), D(z,G) \right), z \in \mathbb{R}^d \right\} \tag{16}$$

Its sample counterpart calculated for two samples $X^n = \{X_1, ..., X_n\}$ from F, and $Y^m = \{Y_1, ..., Y_m\}$ from G is defined as

$$DD(F_n, G_m) = \left\{ \left(D(z,F_n), D(z,G_m) \right), z \in \{X^n \cup Y^m\} \right\} \tag{17}$$

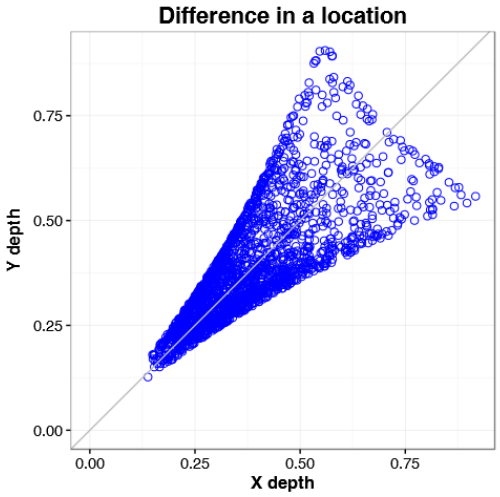

Figure 13: DD-plot, location differences.

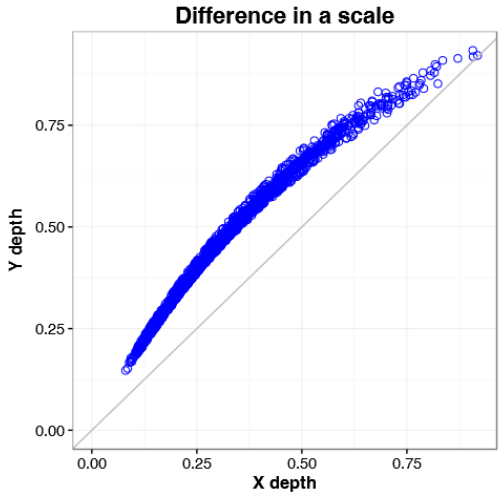

Figure 14: DD-plot, scatter differences.

A detailed presentation of the DD-plot can be found in Liu *et al.* (1999). Fig. 13 presents DD-plot with a heart-shaped pattern in case of differences in location between two samples, whereas Fig. 14 presents a moon-shaped pattern in case of scale differences between samples. Applications of DD-plot and theoretical properties of statistical procedures using this plot can be found in Li and Liu (2004), Liu and Singh (1995), Jurečkova and Kalina (2012), Zuo and He (2006), Kosiorowski and Zawadzki (2014). In Lange *et al.* (2014) an application of the DD-plot for classification can be found.
Within the **DepthProc** we can use DD-plot in a following way:

```
ddPlot(x, y, scale = FALSE, location = FALSE, name_x = "X",
name_y = "Y", title = "Depth vs.  depth plot", ...)
```

Arguments

x: The first or only data sample for ddPlot.
y: The second data sample. x and y must be of the same dimension.
scale : If TRUE samples are centered using multivariate medians.

and

```
ddMvnorm(x, size = nrow(x), robust = FALSE, alpha = 0.05, title = "ddMvnorm", ...)
```

Arguments

x: The data sample for DD plot.
size: Size of theoretical set.
robust: Logical. Default FALSE. If TRUE, robust measures are used to specify the parameters of theoretical distribution.
alpha:cutoff point for robust measure of covariance.

3.5. Multivariate Wilcoxon test

Having two samples $\mathbf{X}^n$ and $\mathbf{Y}^m$ using any depth function, we can compute depth values in a combined sample $\mathbf{Z}^{n+m} = \mathbf{X}^n \cup \mathbf{Y}^m$, assuming the empirical distribution calculated basing on all observations, or only on observations belonging to one of the samples $\mathbf{X}^n$ or $\mathbf{Y}^m$.

For example if we observe $X_l's$ depths are more likely to cluster tightly around the center of the combined sample, while $Y_l's$ depths are more likely to scatter outlying positions, then we conclude $\mathbf{Y}^m$ was drawn from a distribution with larger scale.

Properties of the DD plot based statistics in the i.i.d setting were studied in Li and Liu (2004). Authors proposed several DD-plot based statistics and presented bootstrap arguments for their consistency and good effectiveness in comparison to Hotelling T^2 and multivariate analogues of Ansari-Bradley and Tukey-Siegel statistics. Asymptotic distributions of depth based multivariate Wilcoxon rank-sum test statistic under the null and general alternative hypotheses were obtained in Zuo and He (2006). Several properties of the depth based rang test involving its unbiasedness was critically discussed in Jurečkova and Kalina (2012).

Basing on DD-plot object, which is available within the **DepthProc** it is possible to calculate several multivariate generalizations of one-dimensional rank and order statistics. These generalizations cover well known **Wilcoxon rang-sum statistic**.

The depth based multivariate Wilcoxon rang sum test is especially useful for the multivariate scale changes detection and was introduced among other in Liu and Singh (1995)

For the samples $\mathbf{X}^m = \{\mathbf{X}_1, ..., \mathbf{X}_m\}$, $\mathbf{Y}^n = \{\mathbf{Y}_1, ..., \mathbf{Y}_n\}$, and a combined sample $\mathbf{Z} = \mathbf{X}^n \cup \mathbf{Y}^m$ the **Wilcoxon statistic** is defined as

$$S = \sum_{i=1}^{m} R_i, \tag{18}$$

where R_i denotes the rang of the i-th observation, $i = 1, ..., m$ in the combined sample $R(\mathbf{x}_l) = \#\left\{\mathbf{z}_j \in \mathbf{Z} : D(\mathbf{z}_j, \mathbf{Z}) \leq D(\mathbf{x}_l, \mathbf{Z})\right\}, l = 1, ..., m.$

The distribution of S is symmetric about $E(S) = 1/2m(m+n+1)$, its variance is $D^2(S) = 1/12\ mn(m+n+1)$. For theoretical properties statistic see Li and Liu (2004) and Zuo and He (2006).
Using DD-plot object it is easy to calculate other multivariate test statistics involving for example **Haga** or **Kamat** tests and apply them for robust monitoring of multivariate time series (see Kosiorowski and Zawadzki (2014)).

```
mWilcoxonTest(x, y, alternative = "two.sided")
```

Arguments

x,y: data matrices or data frames of the same dimension
alternative:
Character string determining the alternative, as in one-dimensional Wilcoxon test
method: Character string determining the depth function. method can be "Projection" (the default), "Mahalanobis", "Euclidean", "Tukey", "LP" or "Local".

EXAMPLE

```
> require(MASS)
> x = mvrnorm(100, c(0,0), diag(2))
> y = mvrnorm(100, c(0,0), diag(2)*1.4)
> mWilcoxonTest(x,y)

Multivariate Wilcoxon test for equality of dispersion
data:  dep_x and dep_y
W = 6034, p-value = 0.01156
alternative hypothesis: true dispersion ratio is not equal to 1
```

3.6. Scale and asymmetry curves

For sample depth function $D(x, Z^n)$, $x \in \mathbb{R}^d$, $d \geq 2$, $Z^n = \{z_1, ..., z_n\} \subset \mathbb{R}^d$, $D_\alpha(Z^n)$ denoting $\alpha-$ central region, we can define **the scale curve** (see Fig. 15)

$$SC(\alpha) = (\alpha, vol(D_\alpha(Z^n)) \subset \mathbb{R}^2, \quad for \ \ \alpha \in [0,1], \tag{19}$$

and **the asymmetry curve** (see Fig. 16)

$$AC(\alpha) = \left(\alpha, \left\| c^{-1}(\{\bar{z} - med | D_\alpha(Z^n)\}) \right\|\right) \subset \mathbb{R}^2, \quad for \ \ \alpha \in [0,1] \tag{20}$$

being nonparametric scale and asymmetry functional correspondingly, where $c-$denotes constant, $\bar{z}-$denotes mean vector, denotes multivariate median induced by depth function and $vol-$ denotes a volume. Further information on the scale curve and the asymmetry curve can be found in Liu *et al.* (1999), Serfling and Wang (2006), Serfling (2003), Serfling (2006).

```
scaleCurve(x, y = NULL, alpha = seq(0, 1, 0.01), method = "Projection",
name = "X", name_y = "Y", title = "Scale Curve", ...)
```

Arguments
x: a matrix consisting data.
y: additional data matrix.
alpha: a vector of central regions indices.
method: character string which determines the depth function used, method can be "Projection" (the default), "Mahalanobis", "Euclidean", "Tukey" or "LP".

```
asymmetryCurve(x, y = NULL, alpha = seq(0, 1, 0.01), method = "Projection",
movingmedian = FALSE, name = "X", name_y = "Y", ...) Arguments
```

movingmedian: Logical. For default FALSE only one depth median is used to compute asymmetry norm. If TRUE – for every central area, a new depth median will be used - this approach needs much more computation time.

EXAMPLE

```
> x = mvrnorm(1000, c(0,0),diag(2))
> s1 = scaleCurve(x,name = "Curve 1")
```

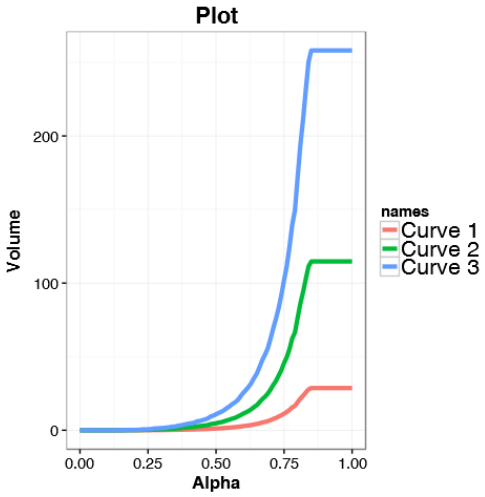

Figure 15: Scale curves.

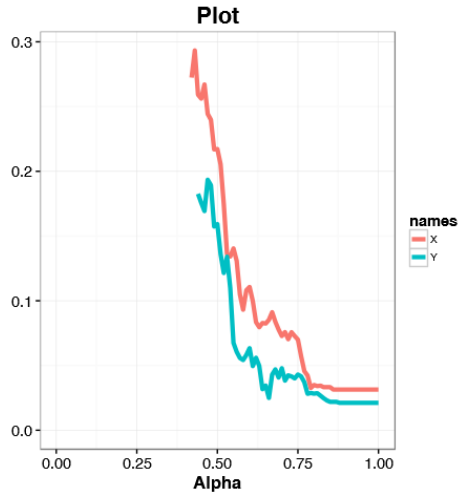

Figure 16: Asymmetry curves.

```
> s2 = scaleCurve(x*2,x*3,name = "Curve 2", name_y = "Curve 3")
> w = getPlot(s1 %+% s2)+ggtitle("Plot")
> w + theme(text = element_text(size = 25))
> xx = mvrnorm(1000, c(0,0),diag(2))
> yy = mvrnorm(1000, c(0,0),diag(2))
> p = asymmetryCurve(xx,yy)
> getPlot(p)+ggtitle("Plot")
```

```
> xx = mvrnorm(1000, c(0,0),diag(2))
> yy = mvrnorm(1000, c(0,0),diag(2))
> p = asymmetryCurve(xx,yy)
> getPlot(p)+ggtitle("Plot")
```

3.7. Simple robust regressions

Within the package two simple (two dimensional) robust regressions are available: **the deepest regression** and **projection depth trimmed regression** – see Fig. 17.

```
deepReg2d(x, y)
trimProjReg2d(x, y, alpha = 0.1)
```

Arguments

x,y: data vectors **alpha**: trimming parameter

EXAMPLE

```
> plot(starsCYG,cex=1.4)
> deepreg = deepReg2d(starsCYG$log.Te, starsCYG$log.light)
> trimreg = trimProjReg2d(starsCYG$log.Te, starsCYG$log.light)
```

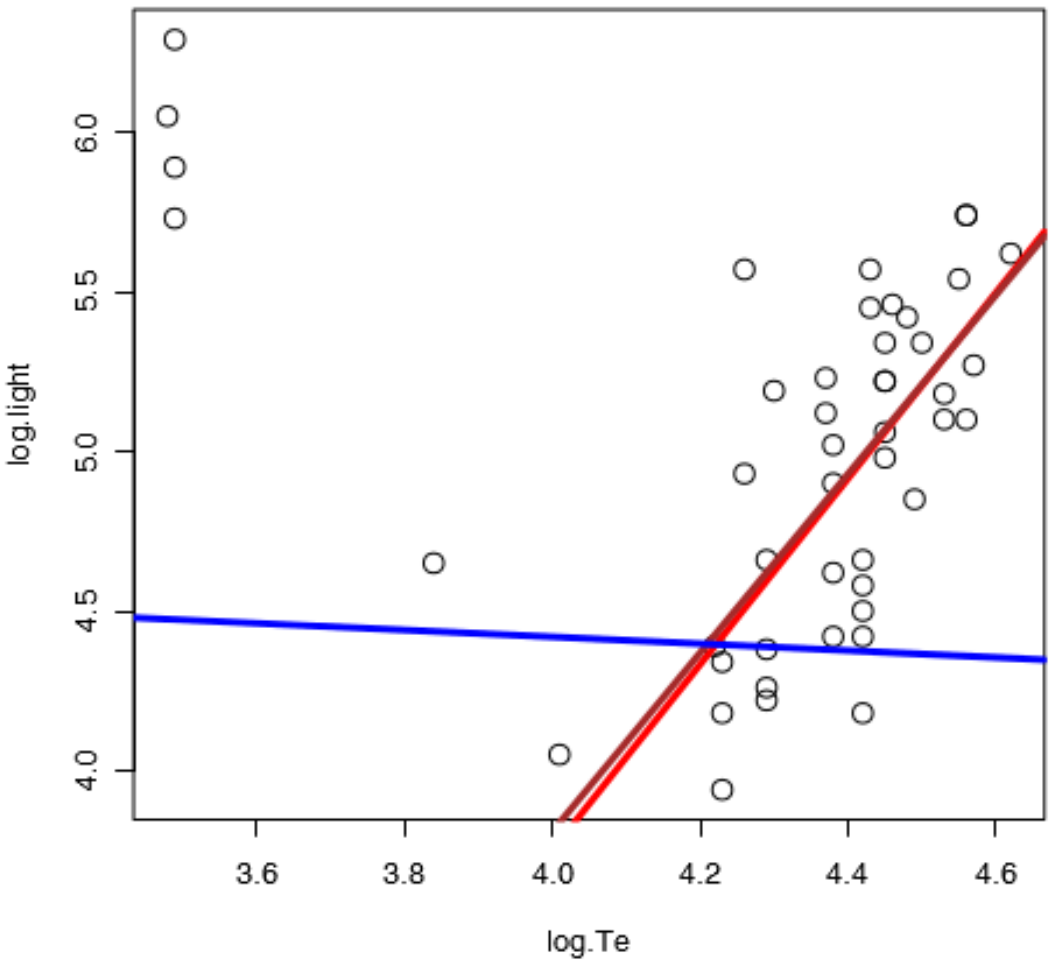

Figure 17: Simple regressions.

```
> least.sq = lm(starsCYG$log.Te~starsCYG$log.light)
> abline(deepreg, lwd = 3, col = "red")
> abline(trimreg, lwd = 3, col = "brown")
> abline(least.sq, lwd = 3, col = "blue")

coefficients:
deepreg@coef
 -7.903043  2.913043
trimreg@coef
-7.403531    2.802837
```

3.8. Weighted estimators of location and scatter

Using depth function one can define a depth-weighted multivariate location and scatter estimators possessing high breakdown points and which for several depths are computationally tractable (see Zuo and Cui (2005)). In case of location, the estimator is defined as

$$L(F) = \int x w_1(D(x,F))dF(x) / w_1(D(x,F))dF(x), \tag{21}$$

Subsequently, a depth-weighted scatter estimator is defined as

$$S(F) = \frac{\int (x - L(F))(x - L(F))^\top w_2(D(x,F))dF(x)}{\int w_2(D(x,F))dF(x)}, \tag{22}$$

where $w_2(\cdot)$ is a suitable weight function that can be different from $w_1(\cdot)$.

The **DepthProc** package offers these estimators in case of computationally feasible weighted L^p depth. Note that$L(\cdot)$ and $S(\cdot)$ include multivariate versions of trimmed means and covariance matrices. Sample counterparts of (20) and (21) take the forms

$$T_{WD}(X^n) = \sum_{i=1}^{n} w(d_i) X_i / \sum_{i=1}^{n} w(d_i), \tag{23}$$

$$DIS(X^n) = \frac{\sum\limits_{i=1}^{n} w(d_i) \left(X_i - T_{WD}(X^n)\right) \left(X_i - T_{WD}(X^n)\right)^T}{\sum\limits_{i=1}^{n} w(d_i)}, \tag{24}$$

where d_i are sample depth weights, $w_1(x) = w_2(x) = a \cdot x + b$, $a, b \in \mathbb{R}$.

Computational complexity of the scatter estimator crucially depend on the complexity of the depth used. For the weighted L^p depth we have $O(d^2n+n^2d)$ complexity and good perspective for its distributed calculation Zuo (2004).

```
CovLP(x, pdim = 1, la = 1, lb = 1)
```

EXAMPLE

```
> require(MASS)
> Sigma1 <- matrix(c(10,3,3,2),2,2)
> X1 = mvrnorm(n= 8500, mu= c(0,0),Sigma1)
> Sigma2 <- matrix(c(10,0,0,2),2,2)
> X2 = mvrnorm(n= 1500, mu= c(-10,6),Sigma2)
> BALLOT<-rbind(X1,X2)
> train <- sample(1:10000, 500)
> data<-BALLOT[train,]
> cov_x = CovLP(data,1,1,1)
> cov_x

Call:
-> Method:  Depth Weighted Estimator
Robust Estimate of Location:
[1]  -1.6980   0.8844
Robust Estimate of Covariance:
          [,1]      [,2]
[1,]  15.249  -2.352
[2,]  -2.352   4.863
```

3.9. Student and L^p binning

Let us recall, that binning is a popular method allowing for faster computation by reducing the continuous sample space to a discrete grid (see Hall and Wand (1996)). It is useful for example in case predictive distribution estimation by means of kernel methods. To bin a window of n points $W_{i,n} = \{X_{i-n+1}, ..., X_i\}$ to a grid $X'_1, ..., X'_m$ we simply assign each sample point X_i to the nearest grid point X'_j. When binning is completed, each grid point X'_j has an associated number c_i, which is the sum of all the points that have been assigned to X'_j. This procedure replaces the data $W_{i,n} = \{X_{i-n+1}, ..., X_i\}$ with the smaller set $W'_{j,m} = \left\{X'_{j-m+1}, ..., X'_j\right\}$. Although simple binning can speed up the computation, it is criticized for a lack of a precise approximate control over the accuracy of the approximation. Robust binning however stresses properties of the majority of the data and decreases the computational complexity of the DSA at the same time.

For a 1D window $W_{i,n}$, let $Z_{i,n-k}$ denote a 2D window created basing on $W_{i,n}$ and consisted of $n-k$ pairs of observations and the k lagged observations $Z_{i,n-k}$=$\{(X_{i-n-k}, X_{i-n+1})\}$, $1 \leq i \leq n-k$. Robust 2D binning of the $Z_{i,n-p}$ is a very useful technique in a context of robust estimation of the predictive distribution of a time series (see Kosiorowski (2013)) or robust monitoring of a data stream (see Kosiorowski and Zawadzki (2014)).

Assume we analyze a data stream $\{X_t\}$ using a moving window of a fixed length n, i.e., $W_{i,n}$ and the derivative window $Z_{i,n-1}$. In a first step we calculate the weighted sample L^p depth for $W_{i,n}$. Next we choose equally spaced grid of points $l_1, ..., l_m$ in this way that $[l_1, l_m]\times[l_1, l_m]$ covers fraction of the β central points of $Z_{i,n-1}$ w.r.t. the calculated L^p depth, i.e., it covers $R^\beta(Z_{i,n-1})$ for certain prefixed threshold $\beta \in (0, 1)$. For both X_t and X_{t-1} we perform a simple binning using following bins: $(-\infty, l_1)$, (l_1, l_2),..., (l_m, ∞).

For robust binning we reject "border" classes and further use only midpoints and binned frequencies for classes (l_1, l_2), (l_2, l_3),..., (l_{m-1}, l_m).

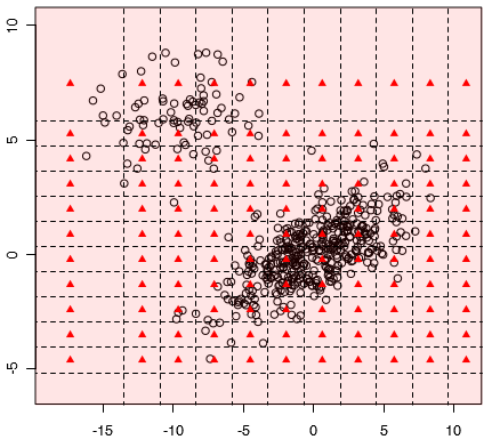

Figure 18: The first step in L^p depth binning.

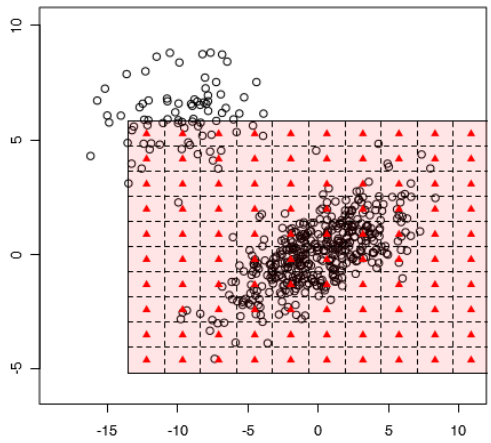

Figure 19: The second step in L^p depth binning.

Figures 18 – 19 present the idea of the simple L^p binning in case of data generated from a mixture of two two-dimensional normal distributions. The midpoints are represented by triangles.

EXAMPLE 1

```
> require(MASS)
> Sigma1 = matrix(c(10,3,3,2),2,2)
> X1 = mvrnorm(n= 8500, mu= c(0,0),Sigma1)
> Sigma2 = matrix(c(10,0,0,2),2,2)
> X2 = mvrnorm(n= 1500, mu= c(-10,6),Sigma2)
> BALLOT = rbind(X1,X2)
> train = sample(1:10000, 500)
> data =BALLOT[train,]
> plot(data)

> b1=binningDepth2D(data, remove_borders = FALSE, nbins = 12, k = 1 )
> b2=binningDepth2D(data, nbins = 12, k = 1,remove_borders = TRUE )
> plot(b1)
> plot(b2)
```

EXAMPLE 2

```
> data(under5.mort)
> data(maesles.imm)
> data2011=cbind(under5.mort[,22],maesles.imm[,22])
> plot(binningDepth2D(data2011, nbins = 8, k = 0.5, remove_borders = TRUE ))
```

4. The package architecture

4.1. Nomenclature conventions

There is no agreed naming convention within R project. In our package we use following coding style:

- *Class* names start with an uppercase letter (e.g. DepthCurve).
- For *methods* and *functions* we use lower camel case convention (e.g. depthTukey)
- All functions related to location-scale depth starts with 'lsd' prefix (e.g. lsdSampleDepthContours).
- Sometimes we depart from these rules whenever to preserve compatibility, with other packages (e.g. CovLP - it is a function from **DepthProc** that follows **rrcov** naming convention).

4.2. Dependencies

Algorithms for depth functions were written in C++, and they are completely independent from R. For matrix operations we use **Armadillo Linear Algebra Library** Sanderson (2010), and **OpenMP** library Board (2013) for parallel computing.

The communication between R and C++ is performed by **RcppArmadillo** package Eddelbuettel and Sanderson (2014).

For plotting we use **base** R graphic (contours plots), **lattice** package Sarkar (2008) (perspective plot), and **ggplot2** Wickham (2009) (other plots). We also uses functions from **rrcov** Todorov and Filzmoser (2009), **np** Hayfield and Racine (2008), **geometry** Barber, Habel, Grasman, Gramacy, Stahel, and Sterratt (2014) packages.

4.3. Parallel computing

By default **DepthProc** uses multi-threading and tries to utilize all available processors. User can control this behaviour with *threads* parameter:

EXAMPLE: Tested on: Intel(R) Core(TM) i5-2500K CPU @ 3.30GHz

```
> x = matrix(rnorm(200000), ncol = 5)
> system.time(depth(x))

user system elapsed
1.484 0.060 0.420
```

EXAMPLE: only one thread (approximately 3 times slower):

```
> system.time(depth(x, threads = 1))
```

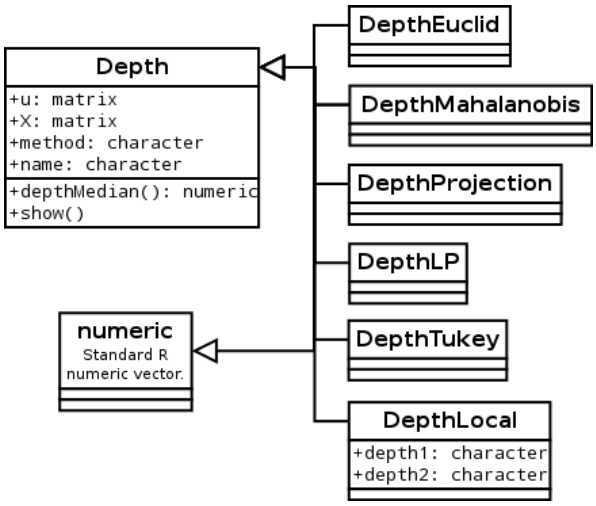

Figure 20: Object structure for classes related to depth functions.

```
user system elapsed
1.368 0.000 1.371
```

EXAMPLE: any value <1 means "use all possible cores"

```
> system.time(depth(x, threads = -10))

user system elapsed
1.472 0.076 0.416
```

4.4. Classes

Below we describe only `Depth`, `DepthCurve`, and `DDPlot` classes in details, because only them have non standard behaviour. Other classes are very simple.

`CovDepthWeighted` is a class for `CovLP` function. It inherits behaviour from `CovRobust` class from **rrcov** package. Description of this class can be found in Todorov and Filzmoser (2009).

4.5. UML diagrams and classes

In this paper we exploit UML class diagrams to describe a behaviour of main **DepthProc** structures. The UML abbreviation stands for *Unified Modelling Language*, a system of notation for describing object oriented programs.

In the UML, class is denoted by a box with three compartments which contain the name, the attributes (slots) and operations (methods) of the class. Each attribute is followed by its type, and each method by its return value. Inheritance relation between classes are depicted by arrowhead pointing to the base class.

4.6. Depth class

Fig. 20 shows an object structure for classes related to depth functions. Each depth class inherits *Depth* and standard *Numeric*. Through inheritance after Numeric these classes are treated as a standard vector, and one can use them with all functions that are appropriate for vectors (e.g. max, min). Depth class is mainly used in internal package operations, but it can be used for extracting depth median without recomputing depth values. This mechanism is show in following example:

EXAMPLE: function for numeric vector

```
> x = matrix(rnorm(1e5), ncol = 2)
> dep = depth(x)

> max(dep)

[1] 0.9860889
```

EXAMPLE: function for raw matrix - all depths must be recomputed:

```
> system.time(dx <- depthMedian(x))

user  system elapsed
1.609   0.072   0.451
```

EXAMPLE: function for Depth class - result is immediate

```
> system.time(dm <- depthMedian(dep))

user  system elapsed
0.000   0.000   0.001
```

In order to check the equality

```
> all.equal(dm, dx)

[1] TRUE
```

4.7. DepthCurve and DDplot classes

The `DepthCurve` is a main class for storing results from `scaleCurve` and the `asymmetryCurve` functions, and describing their behaviour - see Fig. 20. The `DDPlot` stores results from `ddPlot` and `ddMvrnorm` functions.

Both classes `DepthCurve` and `DDPlot` can be converted into **ggplot** object for further appearance modifications via `getPlot()` function.

EXAMPLE:

```
> x = matrix(rnorm(1e2), ncol = 2)
> y = matrix(rnorm(1e2), ncol = 2)

> ddplot =  ddPlot(x,y)
> p = getPlot(ddplot)
```

In order to modify a title

```
> p + ggtitle("X vs Y")
```

```
> scplot = scaleCurve(x,y)
> p = getPlot(scplot)
```

In order to change a color palette:

```
> p + scale_color_brewer(palette = "Set1")
```

Fig. 21 shows class structure for `DepthCurve`. Class `ScaleCurveList` is a container for storing multiple curves for charting them on one plot. It inherits behaviour from standard R list, but it can be also converted into **ggplot** object with `getPlot` method.

We introduced $\%+\%$ operator for combining `DepthCurves` into `DepthCurveList`. This operator is presented in following example:

EXAMPLE

```
> data(under5.mort)
> data(maesles.imm)

> data2011=cbind(under5.mort[,"2011"],maesles.imm[,"2011"])
> data2000=cbind(under5.mort[,"2000"],maesles.imm[,"2000"])
> data1995=cbind(under5.mort[,"1995"],maesles.imm[,"1995"])

> sc2011 = scaleCurve(data2011, name = "2011")
> sc2000 = scaleCurve(data2000, name = "2000")
```

In order to create ScalueCurveList

```
> sclist = sc2000 %+% sc2011
> sclist
```

In order to add another Curve

```
> sc1995 = scaleCurve(data1995, name = "1995")
> sclist %+% sc1995
```

EXAMPLE

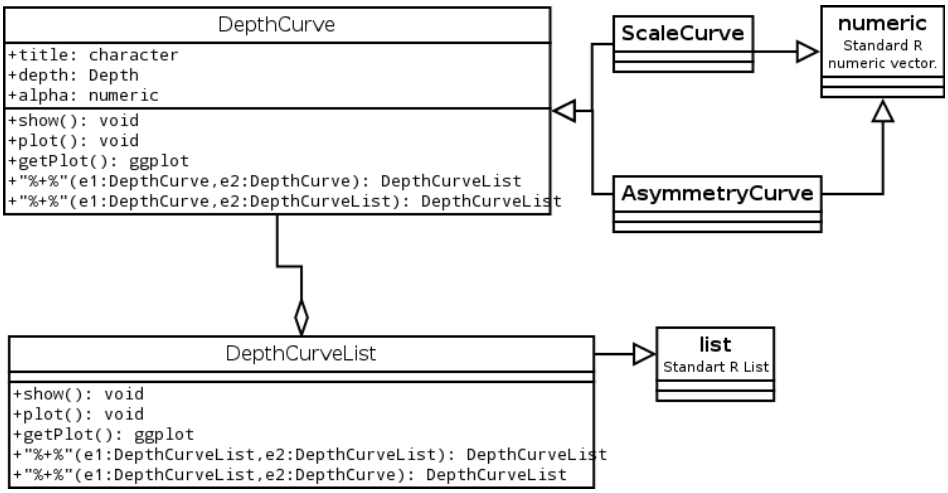

Figure 21: Class structure for `DepthCurve`.

```
> n = 200
> mat_list = replicate(n,matrix(rnorm(200),ncol = 2),simplify = FALSE)
> scurves = lapply(mat_list, scaleCurve)
> scurves = Reduce("%+%",scurves)
> p = getPlot(scurves)
> p + theme(legend.position="none") +
> scale_color_manual(values = rep("black",n))
```

5. Empirical example

For illustrating usefulness of the **DepthProc** package in a socio-economic researches, let us consider an issue of a nonparametric evaluation of the *Fourth Millennium Development Goal* of The United Nations (4MG). Main aim of the goal was reducing by two–thirds, between 1990 – 2015, the under five months child mortality. Using selected multivariate techniques which are available within our **DepthProc** package we answer **a question, if during the period 1990 – 2015 differences between developed and developing countries really decreased.**.

In the study we jointly considered following variables:

- **Children under 5 months mortality rate per 1,000 live births** (Y_1)
- **Infant mortality rate (0–1 year) per 1,000 live births** (Y_2)
- **Children 1 year old immunized against measles, percentage** (Y_3)

Data sets were obtained from `http://mdgs.un.org/unsd/mdg/Data.aspx` and are available within the package. Fig. 22 shows weighted L^2 depth contour with locality parameter $\beta = 0.5$ for countries in 1990 considered w.r.t. variables Y_1 and Y_3 whereas Fig. 5 presents the same issue but in 2011. Fig. 24 shows weighted L^2 depth contour with locality parameter $\beta = 0.5$ for countries in 1990 considered w.r.t. variables Y_2 and Y_3 whereas Fig. 5 presents the same

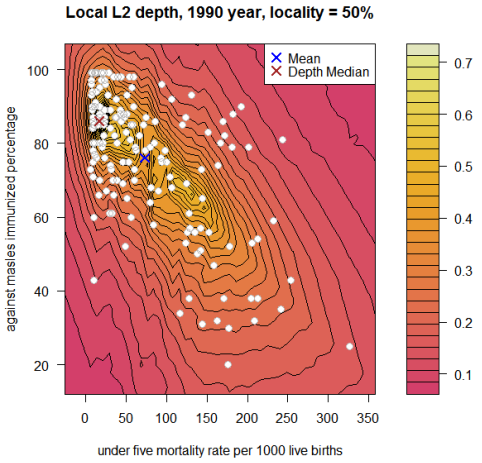

Figure 22: 1990: L^2 depth contour plot Y_1 vs. Y_3

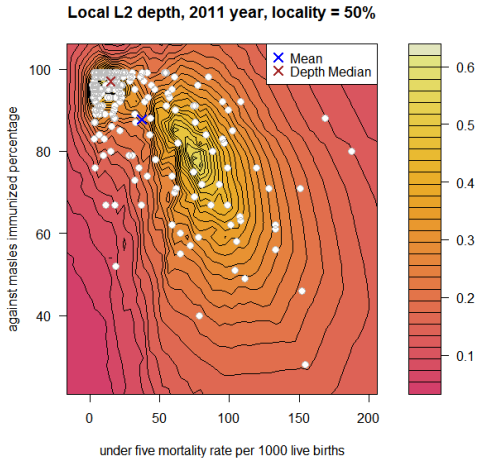

Figure 23: 2011: L^2 depth contour plot Y_1 vs. Y_3

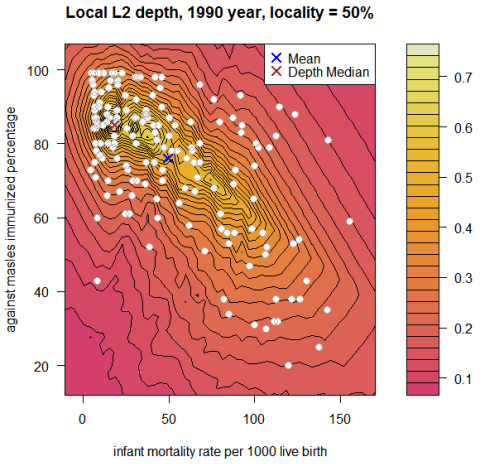

Figure 24: 1990: L^2 depth contour plot Y_2 vs. Y_3

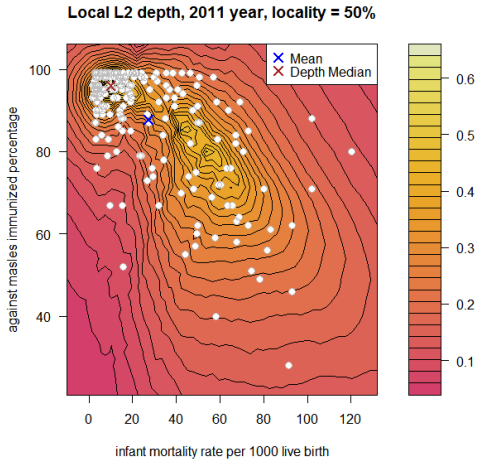

Figure 25: 2011: L^2 depth contour plot Y_2 vs. Y_3

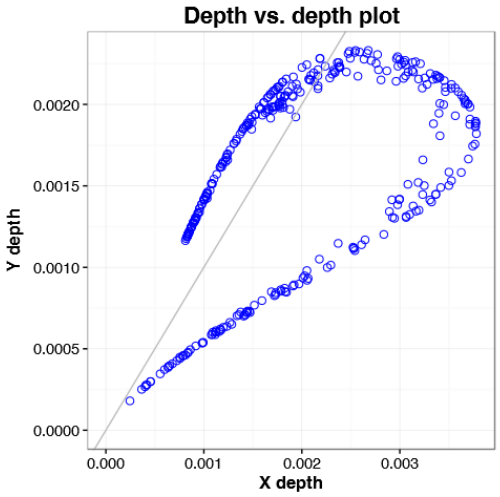

Figure 26: DD plot for inspecting location differences.

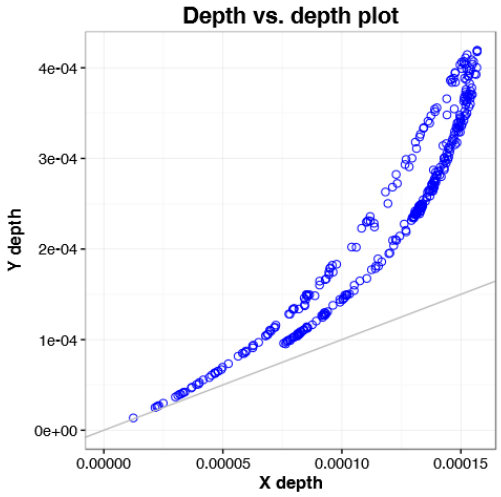

Figure 27: DD plot for inspecting scale differences.

issue but in 2011. Although we can notice a socio-economic development between 1990 and 2011 – the clusters of developed and developing countries are still evident in 2011 as they were in 1990. For assessing changes in location of the centers and scatters of the data between 1990 and 2011 we calculated L^2 **medians** and L^2 **weighted covariance matrices for** (Y_1, Y_2, Y_3) which are presented below

MED(1990): (73.7; 55.2; 78.0)
MED(1995): (59.7; 45.7; 76.0)
MED(2000): (53.7; 42.0; 85.0)
MED(2005): (40.2; 32.6; 86.0)
MED(2010): (33.6; 27.8; 89.0)

$$COV_{L^2}(1990) = \begin{pmatrix} 2420.8 & 1453.9 & -396.3 \\ 1453.9 & 903.4 & -238.6 \\ -396.3 & -238.6 & 228.3 \end{pmatrix}$$

$$COV_{L^2}(2010) = \begin{pmatrix} 738.5 & 493.9 & -158.5 \\ 493.9 & 337.7 & -104.9 \\ -158.5 & -104.9 & 121.2 \end{pmatrix}$$

Fig. 26 presents DD-plot for inspecting location changes between 1990 and 2011 for countries considered w.r.t. variables Y_1, Y_2, Y_3 and Fig. 27 presents DD-plot for inspecting scale changes for the same data. We performed multivariate Wilcoxon test (using L^2 depth) for scale change detection for (Y_1, Y_2, Y_3) in 1990 and in 2011 induced by projection depth and obtained: W=21150 and p-value=0.0046. We can conclude therefore that both the scale and the location changed.

Fig. 28 presents scale curves for the countries considered in the period 1990–2011 jointly w.r.t. all variables whereas Fig. 29 presents Student depth contour plots for variable Y_1 in 1990–2011. **The results of the analysis lead us to following conclusions:**

1. There are big chances for obtaining the 4MG. In the 2010 year, the decrease in the under five months child mortality was about 40% with robust estimates used.

2. For the considered variables, both multivariate as well as univariate, scatters decreased

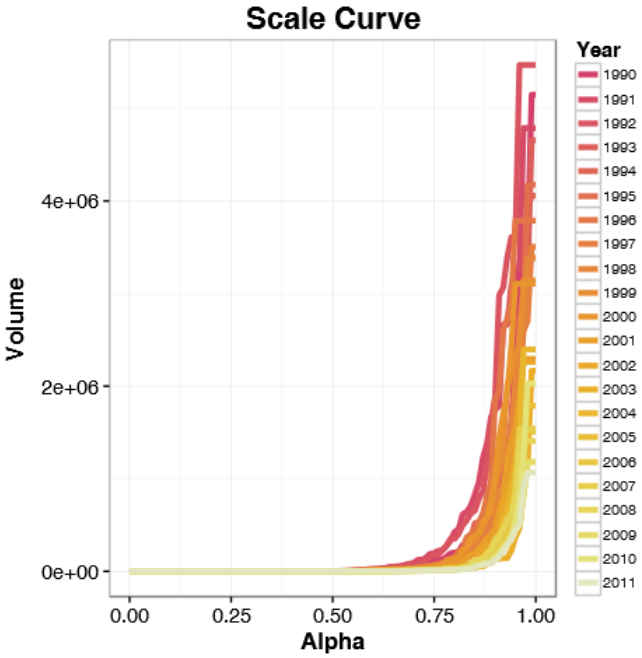

Figure 28: Scale curves for (Y_1, Y_2, Y_3) 1990–2011.

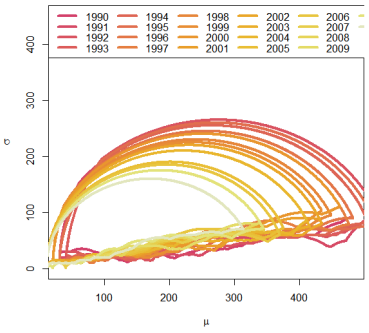

Figure 29: Student depth contour plots - Y_1 in 1990–2011

in 1990–2011.

3. The dispersion between countries considered jointly with respect to variables (Y_1, Y_2, Y_3) significantly decreased in 1990–2011. The clusters of *rich* and *poor* countries are still easily distinguishable however.

4. A comparison of Student depth medians of *Children under 5 months mortality rate per 1,000 live births* in 1990–2011 indicates significant one-dimensional tendency for obtaining the 4MG.

5. Calculated simple deepest regressions for the variables and additional socio-economic variables show clear relations between the 4MG Indicators and with other economic variables representing economic devolvement (e.g., GDP per Capita).

6. The data depth concept offers a complex family of powerful and user-friendly tools for nonparametric and robust analysis of socio-economic multivariate data.

Further considerations related to the issue can be found in Kosiorowska, Kosiorowski, and Zawadzki (2014).

6. Summary

This paper presents **R** package **DepthProc** which offers a selection of multivariate statistical methods originating from the DDC.

Theory of the DDC is still developing by many authors. Recent findings presented in the DDC literature involve among other depths on infinite dimensional spaces, very fast algorithms for approximate depth calculation, new classification rules and new depths on functional spaces. The **DepthProc** package consists of a selection of very powerful but simple and user friendly tools dedicated for a robust economic analysis.

Our plans for a future development of the package concentrate around the concepts of local depth and and depth for functional data. We are going to incorporate these ideas into the Theory of Economics.

Acknowledgement

Daniel Kosiorowski thanks for the polish NCS financial support DEC-011/03/B/HS4/01138.

References

Agostinelli C, Romanazzi M (2013). ***localdepth**: Local Depth*. R package version 0.5-7, URL http://CRAN.R-project.org/package=localdepth.

Barber CB, Habel K, Grasman R, Gramacy RB, Stahel A, Sterratt DC (2014). ***geometry**: Mesh Generation and Surface Tesselation*. R package version 0.3-4, URL http://CRAN.R-project.org/package=geometry.

Board AR (2013). "**OpenMP** Application Program Interface Version 4.0." URL http://www.openmp.org/mp-documents/OpenMP4.0.0.pdf.

Cuesta-Albertos JA, Nito-Reyes A (2008). "The Random Tukey Depth." *Computational Statistics and Data Analysis*, **52**, 4979 – 4988.

Dyckerhoff R (2004). "Data Depths Satisfying the Projection Property." *Allgemeines Statistisches Archiv*, **88**, 163–190.

Eddelbuettel D, Sanderson C (2014). "**RcppArmadillo**: Accelerating R with High-Performance C++ Linear Algebra." *Computational Statistics and Data Analysis*, **71**, 1054–1063. URL http://dx.doi.org/10.1016/j.csda.2013.02.005.

Febrero-Bande M, de la Fuente MO (2012). "Statistical Computing in Functional Data Analysis: The **R** Package **fda.usc**." *Journal of Statistical Software*, **51**(4), 1–28. ISSN 1548-7660. URL http://www.jstatsoft.org/v51/i04.

Genest G, Masse JC, Plante JF (2012). ***depth**: Depth Functions Tools for Multivariate Analysis*. R package version 2.0-0, URL http://CRAN.R-project.org/package=depth.

Hall P, Wand MP (1996). "On the Accuracy of Binned Kernel Density Estimators." *Journal of Multivariate Analysis*, **56**(2), 165–184.

Hayfield T, Racine JS (2008). "Nonparametric Econometrics: The **np** Package." *Journal of Statistical Software*, **27**(5). URL http://www.jstatsoft.org/v27/i05/.

Jurečkova J, Kalina J (2012). "Nonparametric Multivariate Rank Tests and Their Unbiasedness." *Bernoulli*, **18**(1), 229–251.

Kong L, Zuo Y (2010). "Smooth Depth Contours Characterize the Underlying Distribution." *Journal of Multivariate Analysis*, **101**, 2222–2226.

Kosiorowska E, Kosiorowski D, Zawadzki Z (2014). "Evaluation of the Fourth Millenium Developement Goal Realisation Using Multivariate Nonparametric Depth Tools Offered by DepthProc R Package." *submitted*.

Kosiorowski D (2013). "Depth Based Methods for Estimation a Conditional Distribution Function in Data Streams." In *International Statistical Institute Congress Hong Kong 2013 Proceedings*. ISI.

Kosiorowski D (2014). "Robust Estimation of a Data Stream Predictive Distribution Using L^p Depth Binning." *submitted*.

Kosiorowski D, Zawadzki Z (2014). "Multivariate Rank Tests in Economic Data Stream Monitoring." *submitted*.

Lange T, Mosler K, Mozharovskyi P (2014). "Fast Nonparametric Classification Based on Data Depth." *Statistical Papers*, **55**(1), 49–69.

Li J, Liu RY (2004). "New Nonparametric Tests of Multivariate Locations and Scales Using Data Depth." *Statistical Science*, **19**(4), 686–696.

Liu RY, Parelius JM, Singh K (1999). "Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference (with discussion)." *The Annals of Statistics*, **27**, 783–858.

Liu RY, Singh K (1995). "A Quality Index Based on Data Depth and Multivariate Rank Tests." *Journnal of American Statistical Association*, **88**, 252–260.

Lopez-Pintado S, Romo J (2009). "On the Concept of Depth for Functional Data." *Journal of the American Statistical Association*, **104**(486), 718–734.

Lopez-Pintado S, Torrente A (2013). ***depthTools**: Depth Tools Package*. R package version 0.4, URL http://CRAN.R-project.org/package=depthTools.

Maronna RA, Martin RD, Yohai VJ (2006). *Robust Statistics - Theory and Methods*. John Wiley & Sons, Chichester.

Mizera I (2002). "On Depth and Depth Poins: a Calculus." *The Annals of Statistics*, **30**, 1681–1736.

Mizera I, Müller CH (2004). "Location-scale Depth (with discussion)." *Journal of the American Statistical Association*, **99**, 949–966.

Mosler K (2013). "Depth Statistics." In C Becker, R Fried, S Kuhnt (eds.), *Robustness and Complex Data Structures, Festschrift in Honour of Ursula Gather*, pp. 17–34. Springer-Verlag.

Müller C (2003). ***lsdepth**: R-package for Calculating the Student Location-Scale Depth and the Student Median*. URL https://www.statistik.tu-dortmund.de/1253.html.

Paindavaine D, Van Bever G (2012). "Nonparametrically Consistent Depth-Based Classifiers."

Paindavaine D, Van Bever G (2013). "From Depth to Local Depth: a Focus on Centrality." *Journal of the American Statistical Association*, **105**, 1105–1119.

R Core Team (2013). ***R**: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ramsay JO, Hooker G, Graves S (2009). *Functional data analysis with R and Matlab*. Springer-Verlag, New-York.

Rousseeuw PJ, Hubert M (2004). "Regression Depth." *Journal of the American Statistical Association*, **94**, 388–433.

Sanderson C (2010). "**Armadillo**: An Open Source C++ Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments."

Sarkar D (2008). ***Lattice**: Multivariate Data Visualization with R*. Springer-Verlag, New York. ISBN 978-0-387-75968-5, URL http://lmdvr.r-forge.r-project.org.

Serfling R (2003). "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." In RY Liu, R Serfling, DL Souvaine (eds.), *Series in Discrete Mathematics and Theoretical Computer Science*, volume 72, pp. 1–15. AMS.

Serfling R, Wang J (2006). "On Scale Curves for Nonparametric Description of Dispersion." In RY Liu, R Serfling, DL Souvaine (eds.), *Series in Discrete Mathematics and Theoretical Computer Science vol. 72*, pp. 37–48. AMS.

Shao W, Zuo Y (2012). "Simulated Annealing for Higher Dimensional Projection Depth." *Computational Statistics and Data Analysis*, **56**, 4026–4036.

Todorov V, Filzmoser P (2009). "An Object-Oriented Framework for Robust Multivariate Analysis." *Journal of Statistical Software*, **32**(3), 1–47. URL http://www.jstatsoft.org/v32/i03/.

Van Aelst S, Rousseeuw PJ (2000). "Robustness Properties of Deepest Regression." *Journal of Multivariate Analysis*, **73**, 82–106.

Wickham H (2009). ***ggplot2**: Elegant Graphics for Data Analysis*. Springer-Verlag, New York. ISBN 978-0-387-98140-6. URL http://had.co.nz/ggplot2/book.

Zuo Y (2003). "Projection Based Depth Functions and Associated Medians." *The Annals of Statistics*, **31**(5), 1460–1490.

Zuo Y (2004). "Robustness of Weighted L^p Depth and L^p Median." *Allgemaines Statistisches Archiv*, **88**, 215–234.

Zuo Y, Cui H (2005). "Depth Weighted Scatter Estimators." *The Annals of Statistics*, **33**(1), 381 – 413.

Zuo Y, He X (2006). "On the Limiting Distributions of Multivariate Depth-Based Rank Sum Statistics and Related Tests." *The Annals of Statistics*, **34**, 2879–2896.

Zuo Y, Serfling R (2000). "General Notions of Statistical Depth Function." *The Annals of Statistics*, **28**, 461–482.

Affiliation:

Daniel Kosiorowski
Department of Statistics
Faculty of Management
Cracow University of Economics
31-510 Cracow, Poland
E-mail: daniel.kosiorowski@uek.krakow.pl
URL: https://e-uczelnia.uek.krakow.pl/course/view.php?id=137