Depth and depth-based classification with R-package ddalpha

Oleksii Pokotylo*, Pavlo Mozharovskyi**, Rainer Dyckerhoff*

*University of Cologne
**CREST, Ensai

useR!2017

Brussels, July 6, 2017
Contents

Data depth

Depth-based classification

The R-package ddalpha

Summary
Contents

Data depth

Depth-based classification

The R-package ddalpha

Summary
Data depth

Babies with low birth weight

Age, in weeks

Weight, in grams
Data depth

Babies with low birth weight

Age, in weeks

Weight, in grams
Data depth

A data depth measures, how “close” a given point is located to the “center” of a distribution. For \( \mathbf{x} \in \mathbb{R}^d \) and a \( d \)-variate random vector \( \mathbf{X} \) distributed as \( P \in \mathcal{P} \), a data depth is a function

\[
D : \mathbb{R}^d \times \mathcal{P} \to [0, 1], (\mathbf{x}, P) \mapsto D(\mathbf{x} | P)
\]

that is affine invariant, vanishing at infinity, decreasing from deepest point, quasiconcave (upper semicontinuous) in \( \mathbf{x} \).

John W. Tukey (1975) — “Mathematics and the picturing of data”

Tukey depth of \( \mathbf{x} \in \mathbb{R}^d \) w.r.t. a \( d \)-variate random vector \( \mathbf{X} \) distributed as \( P \) is defined as the smallest probability mass of a closed halfspace containing \( \mathbf{x} \):

\[
D^{\text{Tukey}}(\mathbf{x} | \mathbf{X}) = \inf \{ P(H) : H \text{ is a closed halfspace, } \mathbf{x} \in H \}.
\]
Tukey depth
Babies with low birth weight

Tukey depth
Tukey depth

Babies with low birth weight

47 / 161

Weight, in grams

Age, in weeks
Tukey depth

Babies with low birth weight

Age, in weeks

Weight, in grams

26 / 161
Tukey depth

Babies with low birth weight

Age, in weeks

Weight, in grams

41 / 161
Tukey depth

Babies with low birth weight

Age, in weeks

Weight, in grams

49 / 161
Tukey depth

Babies with low birth weight

Weight, in grams

Age, in weeks

114 / 161
Tukey depth

Babies with low birth weight

Age, in weeks

Weight, in grams

800 1000 1200 1400

20 25 30 35

135 / 161
Tukey depth

Babies with low birth weight

Weight, in grams

Age, in weeks

13 / 161
Tukey depth

Babies with low birth weight

Age, in weeks

Weight, in grams

800 1000 1200 1400

20 25 30 35

152 / 161

152 / 161
Tukey depth

Babies with low birth weight

Weight, in grams

Age, in weeks

157 / 161
Tukey depth
Tukey depth

Babies with low birth weight

Age, in weeks

Weight, in grams
Tukey depth

Babies with low birth weight

Age, in weeks

Weight, in grams

9 / 161
Tukey depth

Babies with low birth weight

4 / 161

Weight, in grams

Age, in weeks
Tukey depth

Babies with low birth weight

Weight, in grams

Age, in weeks

9 / 161
Tukey depth

Babies with low birth weight

Weight, in grams

Age, in weeks

147 / 161
Tukey depth

Babies with low birth weight

Age, in weeks

Weight, in grams

3 / 161
Tukey depth
Applications of data depth:

- **Multivariate data analysis** (Liu, Parelius, Singh '99);
- **Statistical quality control** (Liu, Singh '93);
- **Clustering** (Jornsten '04; Jeong, Cai, Sullivan, Wang '16);
- **Tests for multivariate location, scale, symmetry** (Liu '92; Dyckerhoff '02; Dyckerhoff, Ley, Paindaveine '15);
- **Outlier detection** (Hubert, Rousseeuw, Segaert '15);
- **Multivariate risk measurement** (Cascos, Mochalov '07);
- **Robust linear programming** (Bazovkin, Mosler '15);
- etc...

- **Supervised classification** (Ghosh, Chaudhuri '05; Mosler, Hoberg '06; Vencalek '11; Li, Cuesta-Albertos, Liu '12; Lange, Mosler, Mozharovskyi '14; Paindaveine, Van Bever '15; Mosler, Mozharovskyi '15, Pokotylo, Mosler '16, ...);
Contents

Data depth

Depth-based classification

The R-package ddalpha

Summary
Supervised classification

- Random pair \((X, Y)\): \(X\) in \(\mathbb{R}^d\), \(Y\) binary.

- \(X\) has conditional distribution \(P_0\) given \(Y = 0\) resp. \(P_1\) given \(Y = 1\); \(\pi_0 = P(Y = 0), \pi_1 = P(Y = 1)\).

- Given a training sample drawn from \(P_0\) and \(P_1\), \(X_0 = \{x_1, \ldots, x_m\}\) and \(X_1 = \{x_{m+1}, \ldots, x_{m+n}\}\).

- Construct a classification rule \(r\): \(\mathbb{R}^d \rightarrow \{0, 1\}, x \mapsto r(x)\), keeping the classification error small:

\[
\mathcal{E}(r) = \pi_0 P_0(r(X) \neq 0) + \pi_1 P_1(r(X) \neq 1).
\]

- Bayes classifier:

\[
r(x) = \max_{i \in \{0, 1\}} \pi_i f_i(x).
\]
Given: $X_0 = \{x_1, \ldots, x_m\}$ from $P_0$ and $X_1 = \{x_{m+1}, \ldots, x_{m+n}\}$ from $P_1$, consider the DD-plot (Li, Cuesta-Albertos, Liu, 2012),

$$Z = \{z_i | z_i = (D(x_i | X_0), D(x_i | X_1)) \}, \ i = 1, \ldots, m + n.$$
Given: $X_0 = \{x_1, \ldots, x_m\}$ from $P_0$ and $X_1 = \{x_{m+1}, \ldots, x_{m+n}\}$ from $P_1$, consider the $DD$-plot (Li, Cuesta-Albertos, Liu, 2012),

$$Z = \{z_i | z_i = (D(x_i | X_0), D(x_i | X_1)) \}, \ i = 1, \ldots, m + n.$$
**DD-plot**

Given: \( X_0 = \{x_1, \ldots, x_m \} \) from \( P_0 \) and \( X_1 = \{x_{m+1}, \ldots, x_{m+n} \} \) from \( P_1 \),
consider the **DD-plot** (Li, Cuesta-Albertos, Liu, 2012),

\[
Z = \{z_i|z_i = \left( D(x_i|X_0), \ D(x_i|X_1) \right), \ i = 1, \ldots, m+n \}.
\]
Pima Indians Diabetes (Subset: \( m + n = 200, \ d = 7 \))
Pima Indians Diabetes: \textit{DD}-Plot
Extend $DD$-plot using 2nd order polynomial and get 5 features.

In this case $Z = \{ z_i \mid z_i = (D(x_i \mid X_0), D(x_i \mid X_1), D(x_i \mid X_0) \cdot D(x_i \mid X_1), D^2(x_i \mid X_0), D^2(x_i \mid X_1), \ i = 1, \ldots, m + n \}.$

<table>
<thead>
<tr>
<th>Object number</th>
<th>$p_1$ [D_{X_0}(x_i)]</th>
<th>$p_2$ [D_{X_1}(x_i)]</th>
<th>$p_3$ [D_{X_0}(x_i) \cdot D_{X_1}(x_i)]</th>
<th>$p_4$ [D^2_{X_0}(x_i)]</th>
<th>$p_5$ [D^2_{X_1}(x_i)]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$D_{X_0}(x_1)$</td>
<td>$D_{X_1}(x_1)$</td>
<td>$D_{X_0}(x_1) \cdot D_{X_1}(x_1)$</td>
<td>$D^2_{X_0}(x_1)$</td>
<td>$D^2_{X_1}(x_1)$</td>
</tr>
<tr>
<td>2</td>
<td>$D_{X_0}(x_2)$</td>
<td>$D_{X_1}(x_2)$</td>
<td>$D_{X_0}(x_2) \cdot D_{X_1}(x_2)$</td>
<td>$D^2_{X_0}(x_2)$</td>
<td>$D^2_{X_1}(x_2)$</td>
</tr>
<tr>
<td>...</td>
<td>$D_{X_0}(x_i)$</td>
<td>$D_{X_1}(x_i)$</td>
<td>$D_{X_0}(x_i) \cdot D_{X_1}(x_i)$</td>
<td>$D^2_{X_0}(x_i)$</td>
<td>$D^2_{X_1}(x_i)$</td>
</tr>
<tr>
<td>$i$</td>
<td>$D_{X_0}(x_i)$</td>
<td>$D_{X_1}(x_i)$</td>
<td>$D_{X_0}(x_i) \cdot D_{X_1}(x_i)$</td>
<td>$D^2_{X_0}(x_i)$</td>
<td>$D^2_{X_1}(x_i)$</td>
</tr>
<tr>
<td>...</td>
<td>$D_{X_0}(x_{m+n})$</td>
<td>$D_{X_1}(x_{m+n})$</td>
<td>$D_{X_0}(x_{m+n}) \cdot D_{X_1}(x_{m+n})$</td>
<td>$D^2_{X_0}(x_{m+n})$</td>
<td>$D^2_{X_1}(x_{m+n})$</td>
</tr>
</tbody>
</table>
$DD_\alpha$-classifier
$DD_\alpha$-classifier

\[ D(\cdot | X_0) \]

\[ D(\cdot | X_1) \]
$DD\alpha$-classifier
$DD_\alpha$-classifier

\[ D(\cdot | X_0) \]

\[ D(\cdot | X_1) \]
$DD_\alpha$-classifier

\[
D(\cdot|X_0) \cdot D(\cdot|X_1)
\]
$DD_\alpha$-classifier
Contents

Data depth

Depth-based classification

The R-package ddalpha

Summary
Depth-based classification

Data depth + Classification

= 

affine-invariante robust non-parametric distribution-free classification

Problems:

▶ lack of implementations;
▶ different languages and interfaces;
▶ different requirements to the format of the input data;
▶ no implementations of depths and $DD$-classifiers under one roof.

We summarize the work of many researchers.
R-package ddalpha is a structured solution
Implemented data depths

Bivariate points

Mahalanobis

Projection depth

Spatial depth
Implemented data depths

- Tukey depth
- Zonoid depth
- Simplicial depth
- Simplicial volume
Implemented data depths: computation time

Time, sec.

Number of points

- zonoid
- halfspace
- Mahalanobis
- spatial projection
- simplicial
- simplicial volume
### Implemented data depths: algorithms

<table>
<thead>
<tr>
<th>Depth</th>
<th>Exact</th>
<th>Approximate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mahalanobis projection</td>
<td>✓</td>
<td>✓ robust(mcd)</td>
</tr>
<tr>
<td>spatial (L₁) halfspace</td>
<td>✓</td>
<td>✓ pp + ✓ Nelder-Mead</td>
</tr>
<tr>
<td>zonoid</td>
<td>✓</td>
<td>✓ pp</td>
</tr>
<tr>
<td>simplicial</td>
<td>✓</td>
<td>✓ part of simplices</td>
</tr>
<tr>
<td>simplicial volume</td>
<td>✓</td>
<td>✓ part of simplices</td>
</tr>
</tbody>
</table>
Contents

Data depth

Depth-based classification

The R-package ddalpha

Summary
Summary of the R-package ddalpha

Package ‘ddalpha’

Type Package
Title Depth-Based Classification and Calculation of Data Depth
Version 1.2.1
Date 2016-10-09
SystemRequirements C++11
Depends stats, utils, graphics, grDevices, MASS, class, robustbase
Imports Rcpp (>= 0.11.0)
LinkingTo BH, Rcpp
Description Contains procedures for depth-based supervised learning, which are entirely non-parametric, in particular the DDelph procedure (Lange, Mosler and Mozharovskyi, 2014). The training data sample is transformed by a statistical depth function to a compact low-dimensional space, where the final classification is done. It also offers an extension to functional data and routines for calculating certain notions of statistical depth functions. 50 multivariate and 5 functional classification problems are included.
License GPL-2
NeedsCompilation yes
Author Oleksii Pokotylo [aut, cre], Pavlo Mozharovskyi [aut], Rainer Dyckerhoff [aut]
Maintainer Oleksii Pokotylo <alexeypokotylo@gmail.com>
Repository CRAN
Date/Publication 2016-10-10 01:48:09

- exact and approximate computation of 7 data depths
- depth-based supervised classification
- supports multivariate and functional data
- outsiders treatment procedures
- built in procedures for statistical inference
- data sets and data generators
- visualization procedures
Thank you for your attention! Questions?


