Depth and depth-based classification with R-package ddalpha

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

*University of Cologne
**CREST, Ensai, Université Bretagne Loire
***Charles University in Prague

Septièmes rencontres R

Rennes, 6 juillet 2018
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

Weight, in grams

Age, in weeks
Data depth

Babies with low birth weight

Age, in weeks

Weight, in grams

800 1000 1200 1400

20 25 30 35
Data depth

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

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

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

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

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

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

Babies with low birth weight

Age, in weeks

Weight, in grams

120 / 161
Tukey depth

Babies with low birth weight

Weight, in grams

Age, in weeks

112 / 161
Tukey depth

Babies with low birth weight

800 1000 1200 1400
20 25 30 35

Age, in weeks

Weight, in grams

47 / 161
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: 20 to 35
- Weight, in grams: 800 to 1400

41 / 161
Tukey depth

Babies with low birth weight

49 / 161

Age, in weeks

Weight, in grams
Tukey depth

Babies with low birth weight

Weight, in grams

Age, in weeks

114 / 161
Tukey depth

Babies with low birth weight

Weight, in grams

Age, in weeks
Tukey depth

Babies with low birth weight

Age, in weeks

Weight, in grams

13 / 161
Tukey depth

Babies with low birth weight

Weight, in grams

Age, in weeks
Tukey depth

Babies with low birth weight

Weight, in grams

Age, in weeks

Babies with low birth weight
Tukey depth

Babies with low birth weight

Weight, in grams

Age, in weeks

157 / 161
Tukey depth
Tukey depth

Babies with low birth weight

Weight, in grams

Age, in weeks

14/161
Tukey depth

Babies with low birth weight

Weight, in grams

Age, in weeks

4 / 161
Tukey depth

Babies with low birth weight

Age, in weeks

Weight, in grams

9 / 161
Tukey depth

Babies with low birth weight

Age, in weeks

Weight, in grams

147 / 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);
- **Missing data imputation** (Mozharovskyi, Josse, Husson ’17);
- 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),

\[
Z = \{ z_i | z_i = \left( D(x_i | X_0), \ D(x_i | X_1) \right), \ i = 1, \ldots, m + n \}.
\]
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),

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

$$Z = \{z_i \mid z_i = \left( D(x_i \mid X_0), \ D(x_i \mid X_1) \right), \ i = 1, \ldots, m + n \}.$$
Pima Indians Diabetes (Subset: $m + n = 200, d = 7$)
Pima Indians Diabetes: \textit{DD}-Plot

![Graph showing depth with respect to classes X and X0]
**DDα-classifier**

Extend **DD-plot using** 2nd order **polynomial** and get **5 features**.

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

<table>
<thead>
<tr>
<th>Object number</th>
<th>( \frac{p_1}{D_{X_0}(x_i)} )</th>
<th>( \frac{p_2}{D_{X_1}(x_i)} )</th>
<th>( \frac{p_3}{D_{X_0}(x_i) \cdot D_{X_1}(x_i)} )</th>
<th>( \frac{p_4}{D^2_{X_0}(x_i)} )</th>
<th>( \frac{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_{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

\[ D(\cdot|X_0) \]

\[ D(\cdot|X_1) \]

\[ \alpha^{(1)} \]
$DD_\alpha$-classifier
$\mathbf{DD}_\alpha$-classifier
$DD_\alpha$-classifier
Depth-based classification

Data depth + Classification

= 

affine-invariant 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

- 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 ($L_1$) halfspace</td>
<td>✓ ✓ ✓</td>
<td>✓ pp + ✓ Nelder-Mead</td>
</tr>
<tr>
<td>zonoid</td>
<td>✓</td>
<td>✓ robust(mcd)</td>
</tr>
<tr>
<td>simplicial</td>
<td>✓</td>
<td>✓ pp</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’

June 23, 2018

Type Package
Title Depth-Based Classification and Calculation of Data Depth
Version 1.3.4
Date 2018-06-22
SystemRequirements C++11
Depends R (>= 2.10), stats, utils, graphics, grDevices, MASS, class, robustbase, sfsmisc, geometry
Imports Repp (>= 0.11.0)
LinkingTo BH, Repp
Description Contains procedures for depth-based supervised learning, which are entirely non-parametric, in particular the DDaflow procedure (Lange, Mosler and Mozharovskyi, 2014 <doi:10.1007/s00362-012-0488-4>). 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], Stanislav Nagy [aut]
Maintainer Oleksii Pokotylo <alexey.pokotylo@gmail.com>
Repository CRAN
Date/Publication 2018-06-23 16:08:17 UTC

- 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?


