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

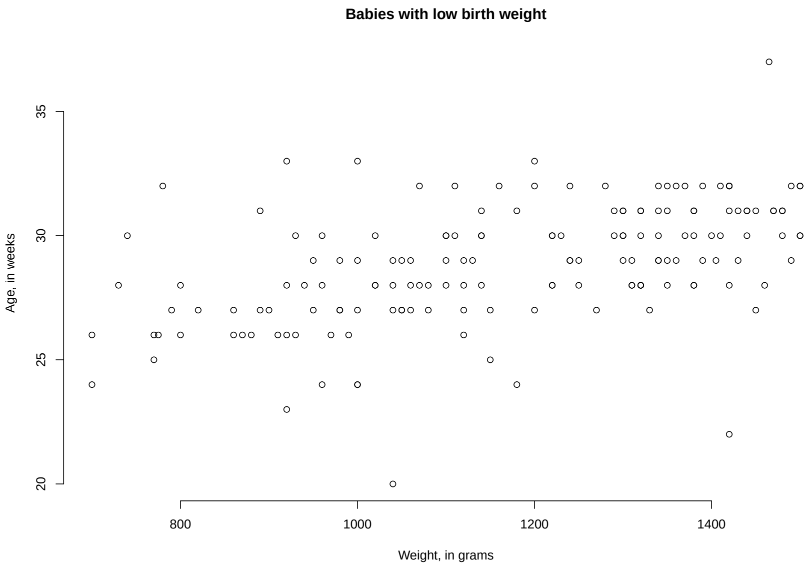
Septièmes rencontres R

Rennes, 6 juillet 2018

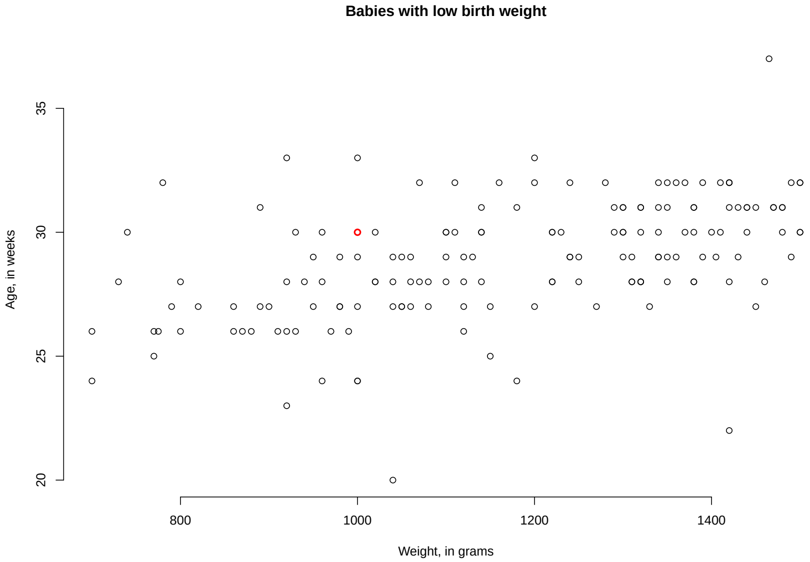
Contents

Contents

Data depth


Babies with low birth weight
Age, in weeks
35
30
25
20
800
1000
1200
1400
Weight, in grams

Data depth



Babies with low birth weight
Age, in weeks
20
25
30
35
800
1000
1200
1400
Weight, in grams

Data depth

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

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

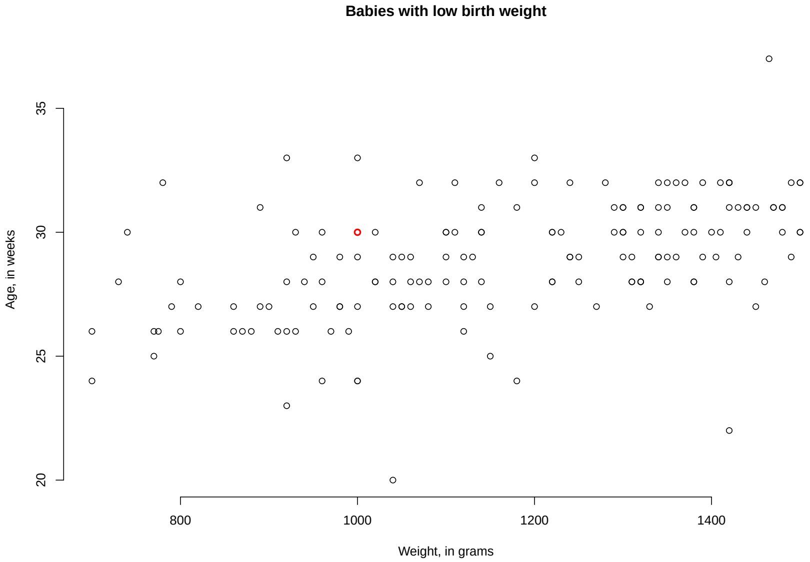
that is **affine invariant**, **vanishing at infinity**, **decreasing** from deepest point, **quasiconcave** (upper semicontinuous) in **x**.

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

Tukey depth of $\boldsymbol{x} \in \mathbb{R}^d$ 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**:

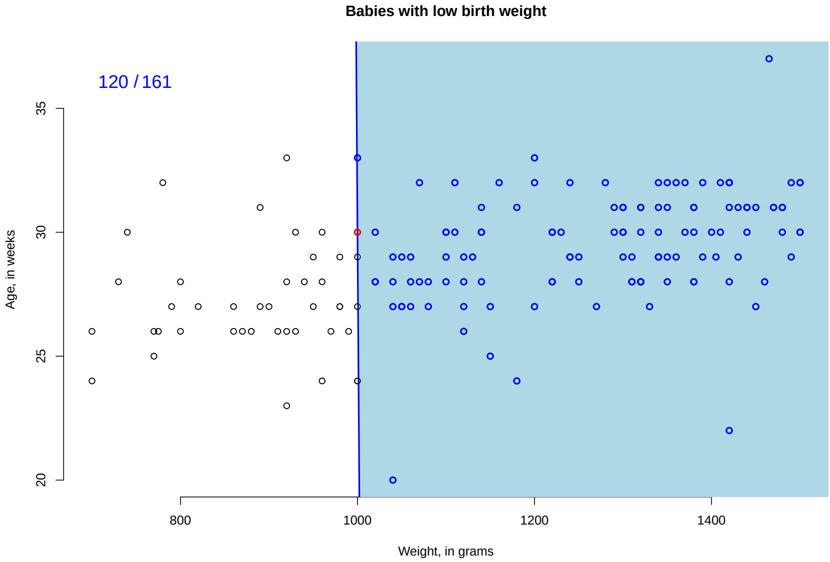
$$D^{Tukey}(\boldsymbol{x}|X) = \inf\{P(H) \,:\, H \text{ is a closed halfspace, } \boldsymbol{x} \in H\}.$$

Tukey depth

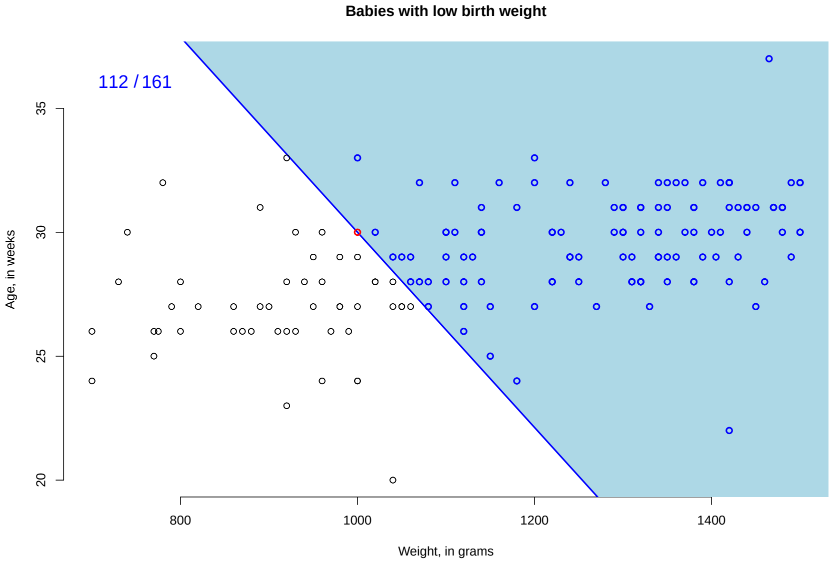


Babies with low birth weight
Age, in weeks
20
25
30
35
800
1000
1200
1400
Weight, in grams

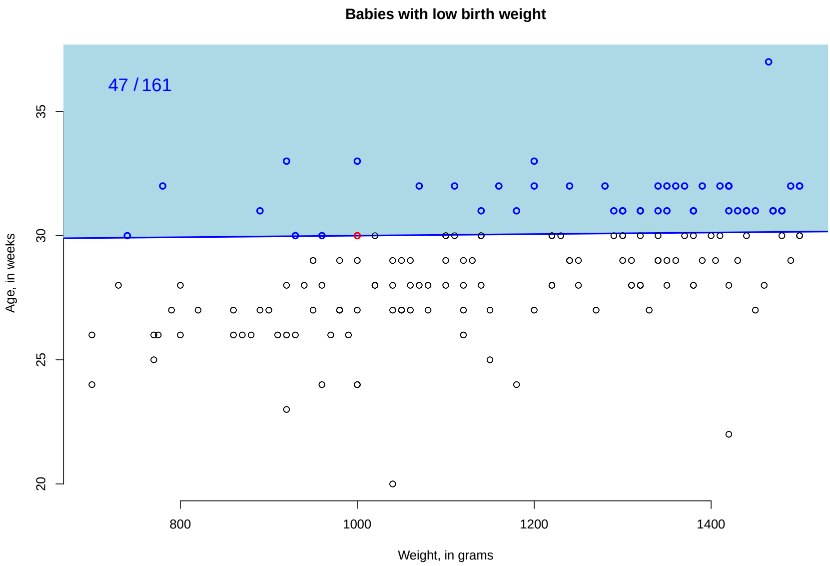
Tukey depth


Babies with low birth weight
120 / 161
Age, in weeks
Weight, in grams
20
25
30
35
800
1000
1200
1400

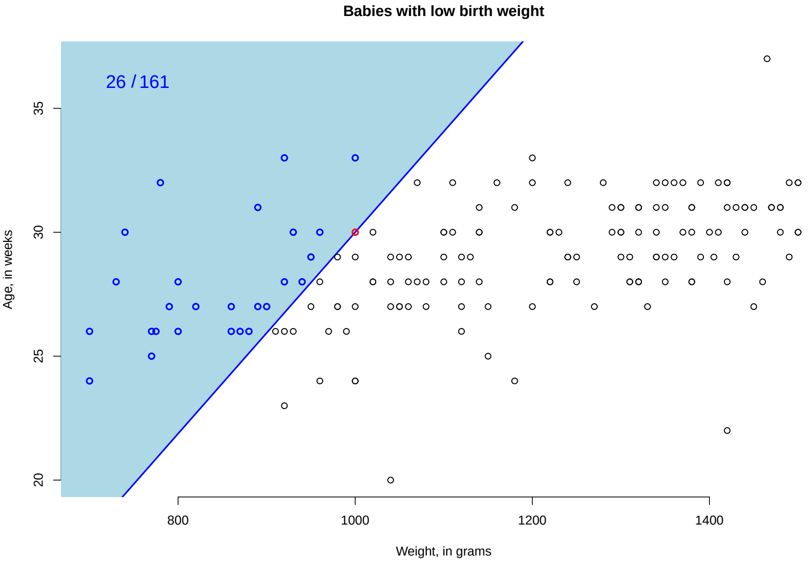
Tukey depth


Babies with low birth weight
112 / 161
Age, in weeks
20
25
30
35
800
1000
1200
1400
Weight, in grams

Tukey depth

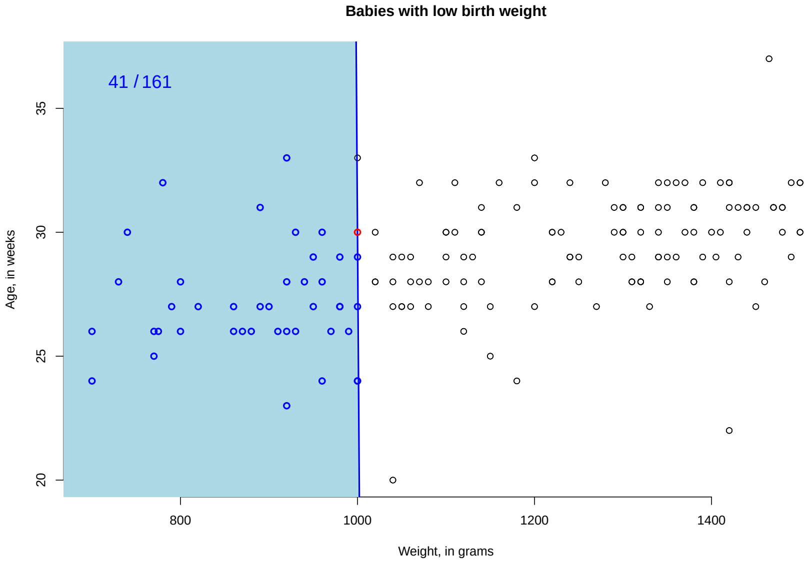

Babies with low birth weight
47 / 161
Age, in weeks
20
25
30
35
800
1000
1200
1400
Weight, in grams

Tukey depth



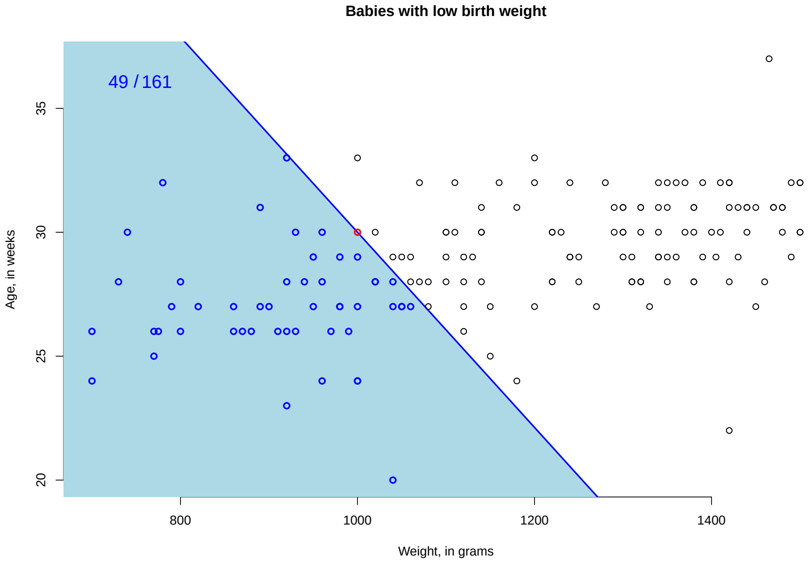
Babies with low birth weight
26 / 161
Age, in weeks
20
25
30
35
800
1000
1200
1400
Weight, in grams

Tukey depth



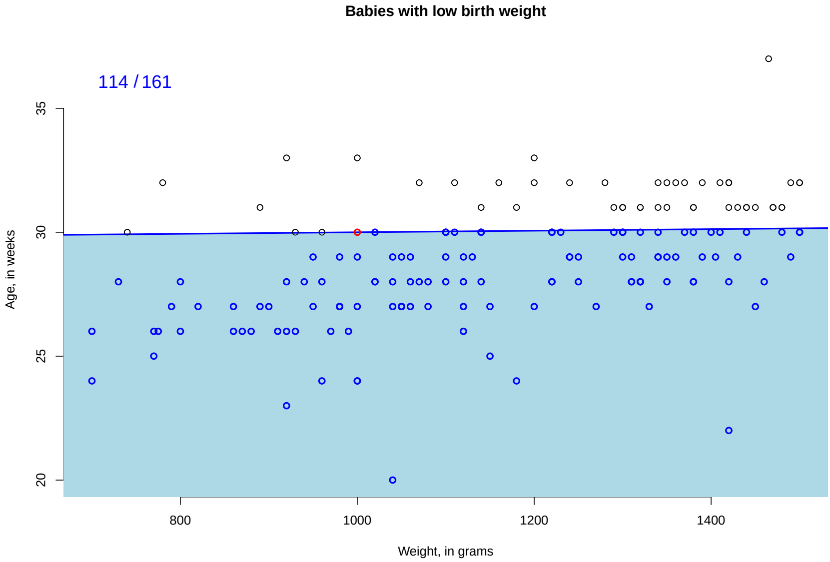
Babies with low birth weight
41 / 161
35
30
25
20
Age, in weeks
800
1000
1200
1400
Weight, in grams

Tukey depth



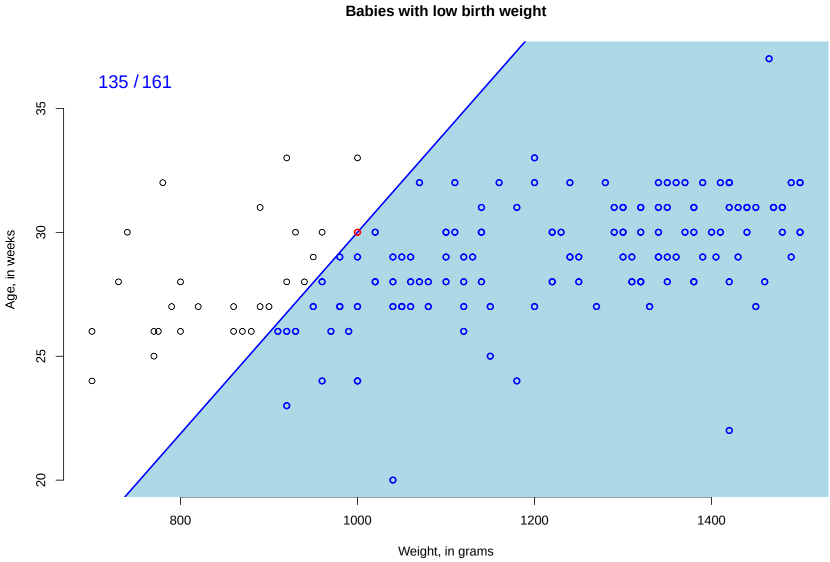
Babies with low birth weight
49 / 161
Age, in weeks
20
25
30
35
800
1000
1200
1400
Weight, in grams

Tukey depth

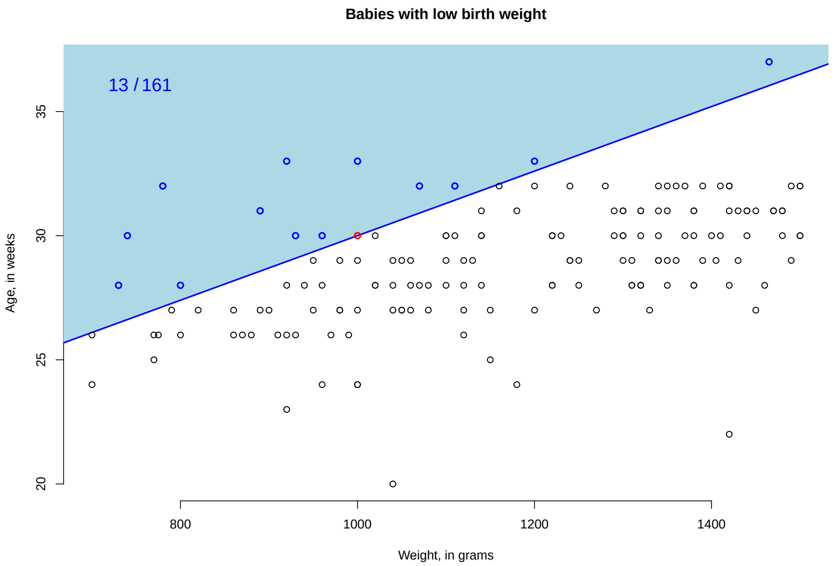


Babies with low birth weight
114 / 161
35
30
25
20
Age, in weeks
800
1000
1200
1400
Weight, in grams

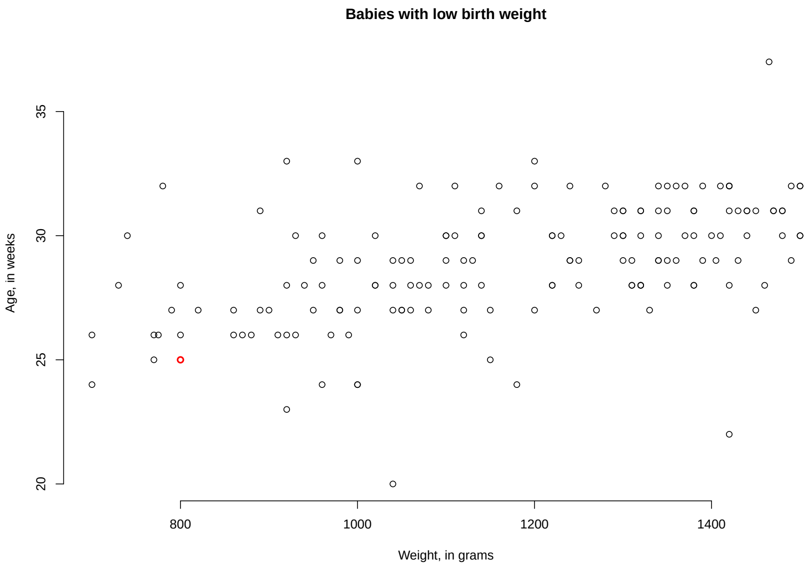
Tukey depth


Babies with low birth weight
135 / 161
Age, in weeks
20
25
30
35
800
1000
1200
1400
Weight, in grams

Tukey depth

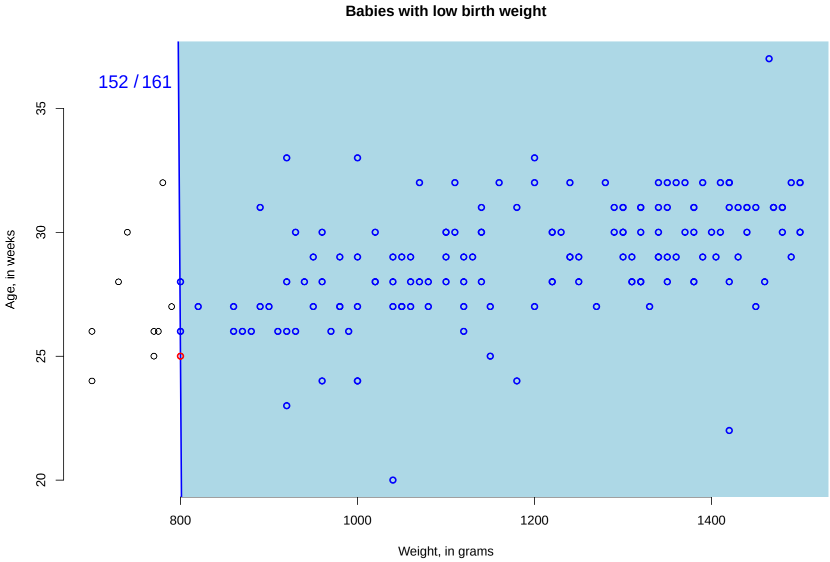

Babies with low birth weight
13 / 161
Age, in weeks
20
25
30
35
800
1000
1200
1400
Weight, in grams

Tukey depth

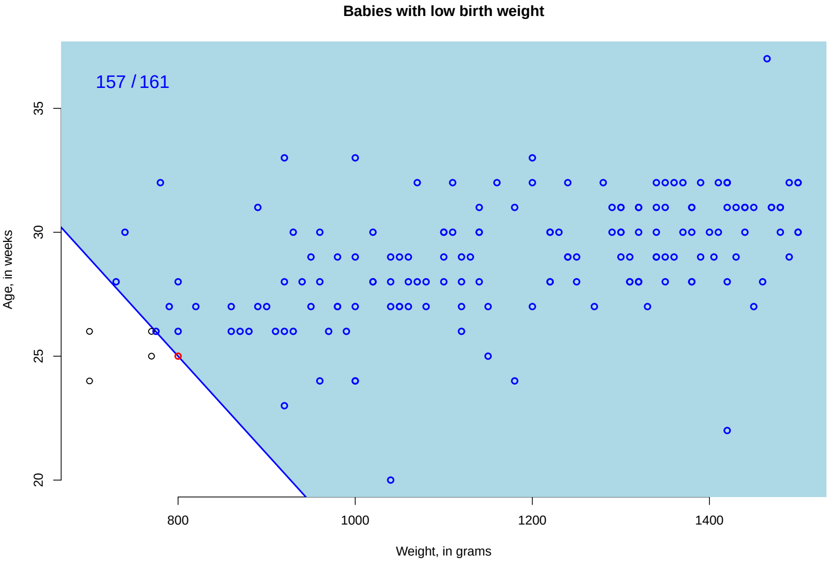


Babies with low birth weight
Age, in weeks
35
30
25
20
800
1000
1200
1400
Weight, in grams

Tukey depth

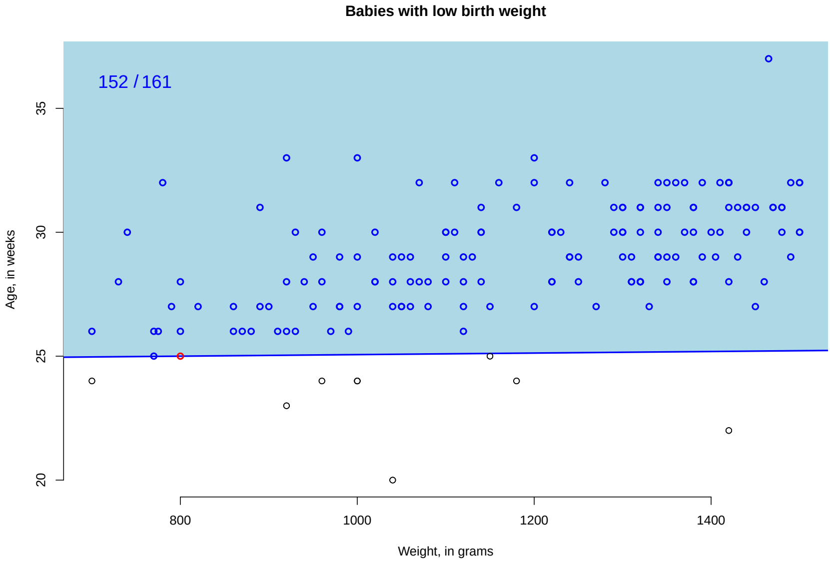

Babies with low birth weight
152 / 161
35
30
25
20
Age, in weeks
800
1000
1200
1400
Weight, in grams

Tukey depth



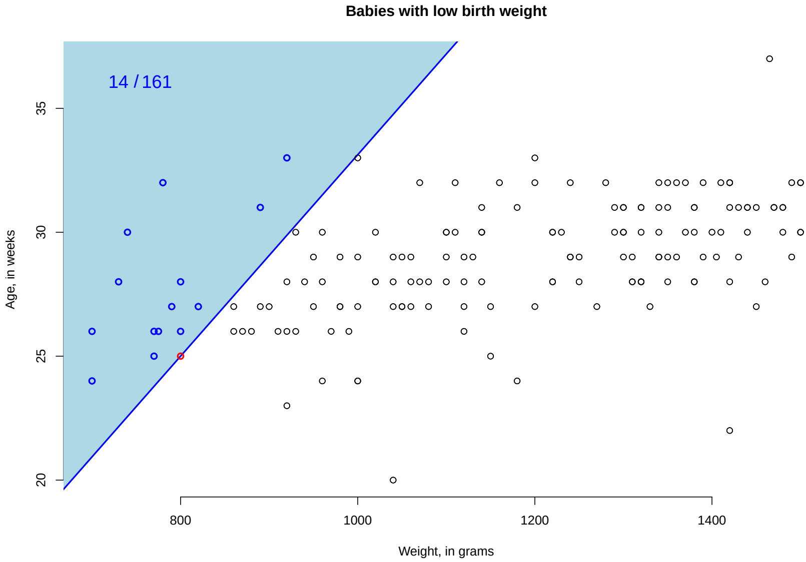
Babies with low birth weight
157 / 161
Age, in weeks
Weight, in grams
35
30
25
20
800
1000
1200
1400

Tukey depth



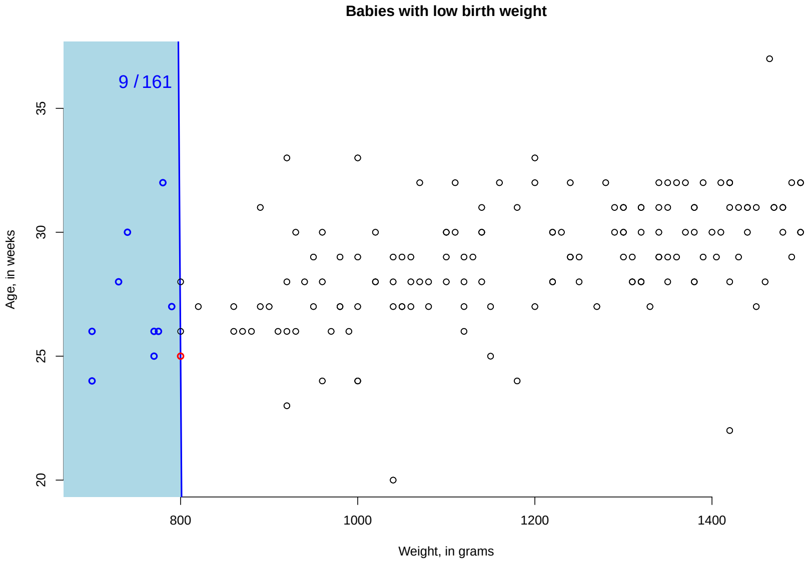
Babies with low birth weight
152 / 161
Age, in weeks
Weight, in grams
20
25
30
35
800
1000
1200
1400

Tukey depth

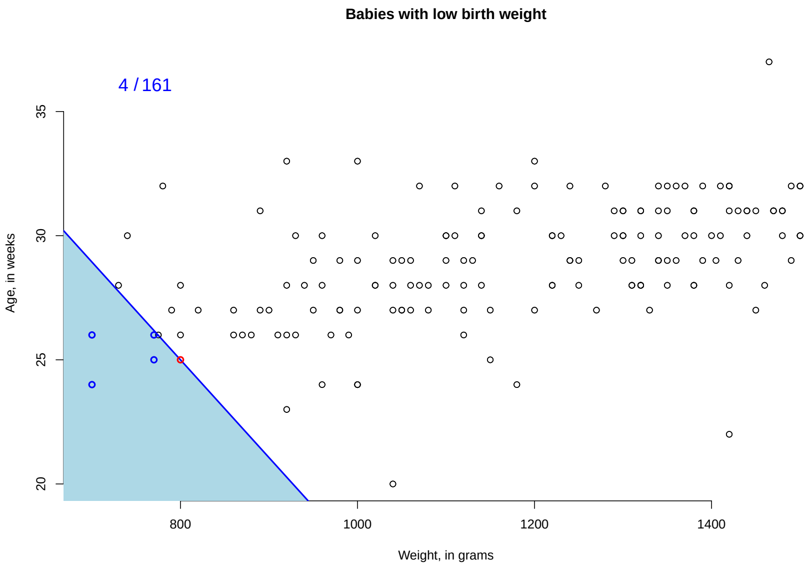


Babies with low birth weight
14 / 161
Age, in weeks
20
25
30
35
800
1000
1200
1400
Weight, in grams

Tukey depth

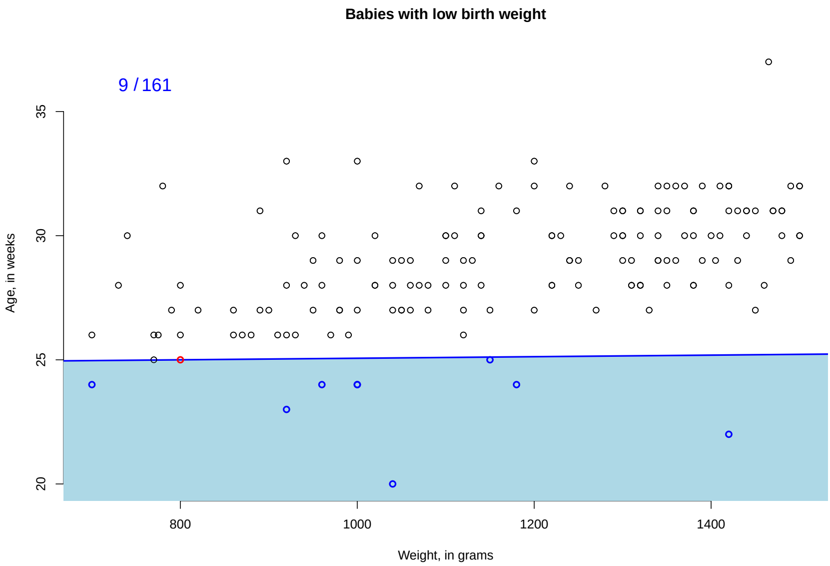

Babies with low birth weight
9 / 161
35
30
25
20
Age, in weeks
800
1000
1200
1400
Weight, in grams

Tukey depth



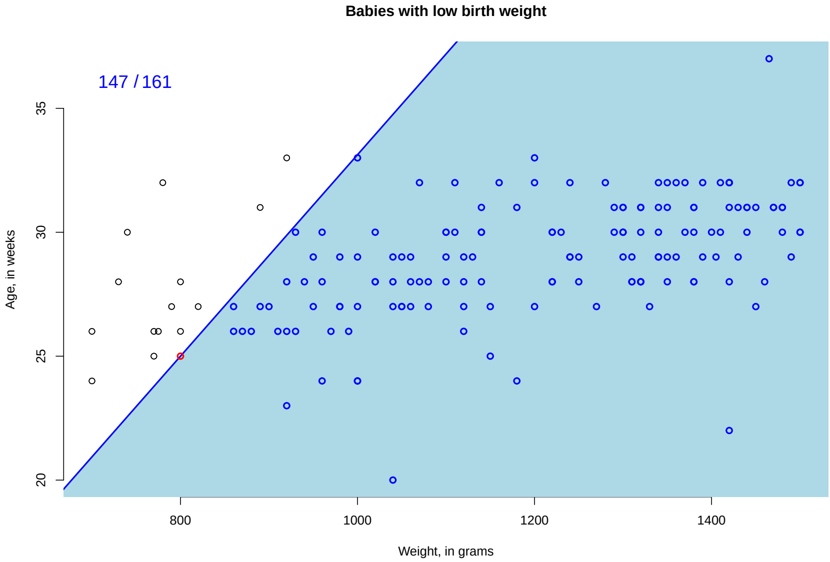
Babies with low birth weight
4 / 161
35
30
25
20
Age, in weeks
800
1000
1200
1400
Weight, in grams

Tukey depth

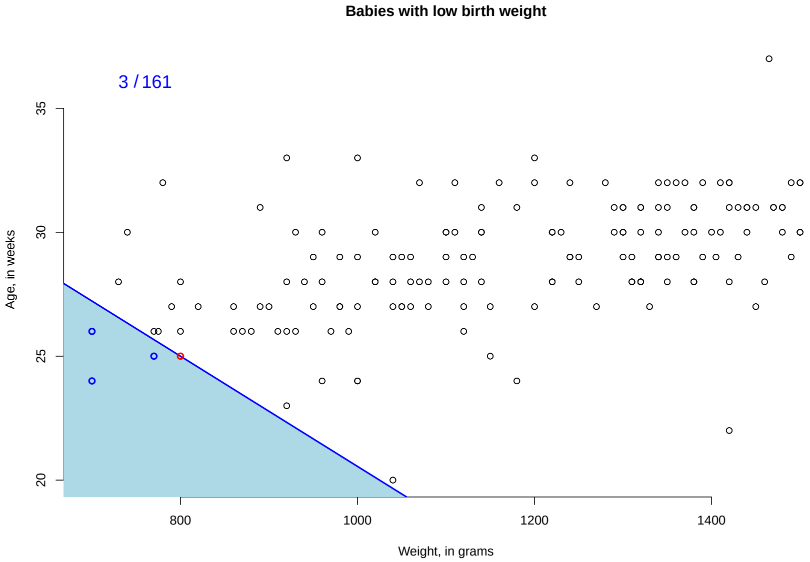


Babies with low birth weight
9 / 161
35
30
25
20
Age, in weeks
800
1000
1200
1400
Weight, in grams

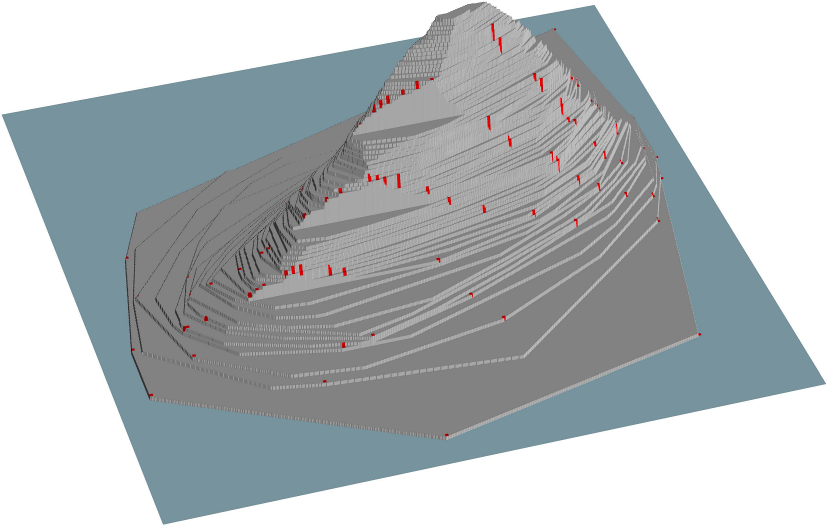
Tukey depth


Babies with low birth weight
147 / 161
Age, in weeks
20
25
30
35
800
1000
1200
1400
Weight, in grams

Tukey depth


Babies with low birth weight
3 / 161
Age, in weeks
35
30
25
20
800
1000
1200
1400
Weight, in grams

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

Supervised classification

- Random pair (X, Y): X in $\mathbf{R}^d$, 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, $X_0 = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$ and $X_1 = \{\mathbf{x}_{m+1}, ..., \mathbf{x}_{m+n}\}$,
- construct a **classification rule** $\boldsymbol{r}$: $\mathbb{R}^d \to \{0, 1\}$, $\mathbf{x} \mapsto \boldsymbol{r}(\mathbf{x})$, keeping the classification error small:

$$\mathcal{E}(\boldsymbol{r}) = \pi_0 P_0\big(\boldsymbol{r}(X) \neq 0\big) + \pi_1 P_1\big(\boldsymbol{r}(X) \neq 1\big).$$

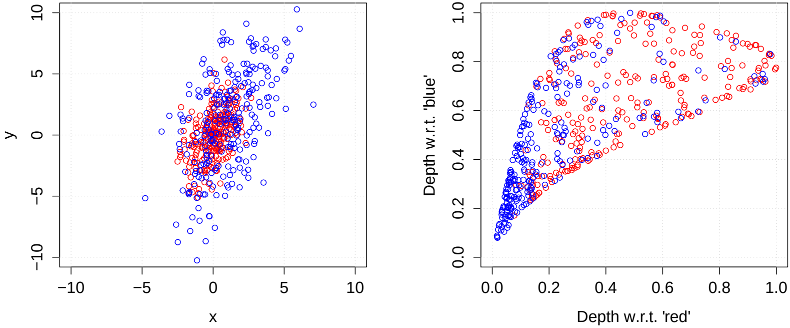
- **Bayes classifier**:

$$\boldsymbol{r}(\mathbf{x}) = \max_{i \in \{0,1\}} \pi_i f_i(\mathbf{x}).$$

DD-plot

Given: $X_0 = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$ from P_0 and $X_1 = \{\mathbf{x}_{m+1}, ..., \mathbf{x}_{m+n}\}$ from P_1, consider the DD-plot (Li, Cuesta-Albertos, Liu, 2012),

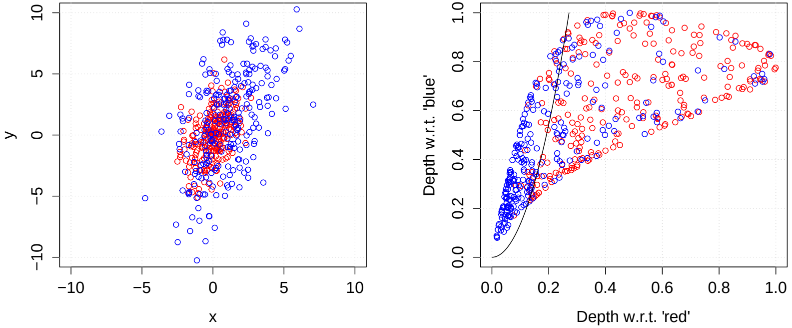
$$Z = \{\mathbf{z}_i | \mathbf{z}_i = \big(\; D(\mathbf{x}_i|X_0), \;\; D(\mathbf{x}_i|X_1) \; \big), \;\; i = 1, ..., m+n\}.$$



DD-plot

Given: $X_0 = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$ from P_0 and $X_1 = \{\mathbf{x}_{m+1}, ..., \mathbf{x}_{m+n}\}$ from P_1, consider the DD-plot (Li, Cuesta-Albertos, Liu, 2012),

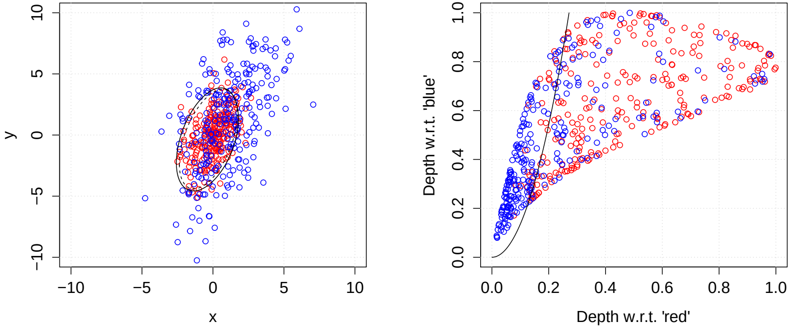
$$Z = \{\mathbf{z}_i | \mathbf{z}_i = \big(\; D(\mathbf{x}_i|X_0), \;\; D(\mathbf{x}_i|X_1) \; \big), \;\; i = 1, ..., m+n\}.$$



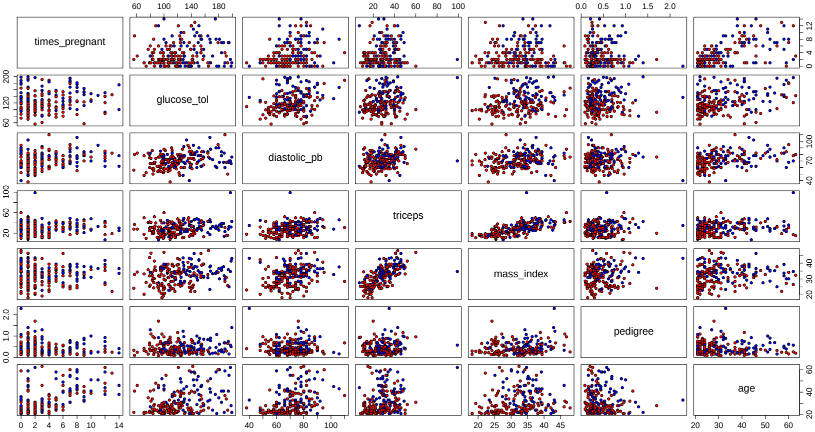
DD-plot

Given: $X_0 = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$ from P_0 and $X_1 = \{\mathbf{x}_{m+1}, ..., \mathbf{x}_{m+n}\}$ from P_1, consider the DD-plot (Li, Cuesta-Albertos, Liu, 2012),

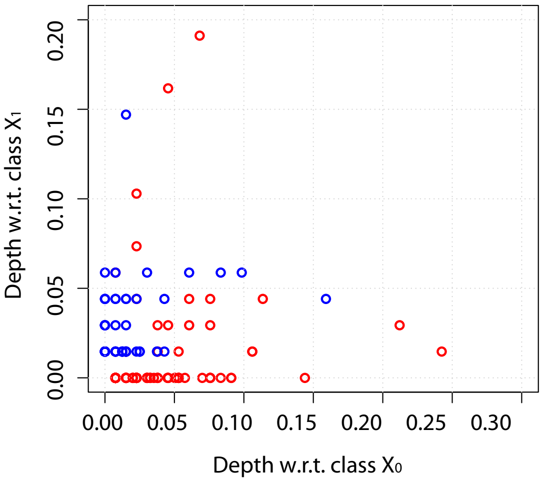
$$Z = \{\mathbf{z}_i | \mathbf{z}_i = \left(\; D(\mathbf{x}_i | X_0), \;\; D(\mathbf{x}_i | X_1) \; \right), \;\; i = 1, ..., m+n\}.$$



Pima Indians Diabetes (Subset: $m + n = 200$, $d = 7$)



Pima Indians Diabetes: *DD*-Plot



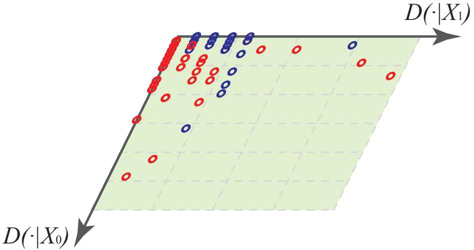
$DD\alpha$-classifier

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

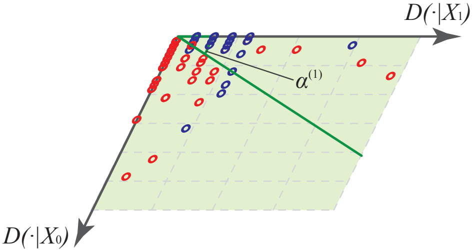
In this case $Z = \{\mathbf{z}_i | \mathbf{z}_i = \big(\; D(\mathbf{x}_i|X_0), \; D(\mathbf{x}_i|X_1),$
$D(\mathbf{x}_i|X_0) \cdot D(\mathbf{x}_i|X_1), \; D^2(\mathbf{x}_i|X_0), \; D^2(\mathbf{x}_i|X_1) \; \big), \; i = 1, ..., m+n\}$.

Object number	Extended properties				
	$\underline{p_1}$ $D_{X_0}(\mathbf{x}_i)$	$\underline{p_2}$ $D_{X_1}(\mathbf{x}_i)$	$\underline{p_3}$ $D_{X_0}(\mathbf{x}_i) \cdot D_{X_1}(\mathbf{x}_i)$	$\underline{p_4}$ $D^2_{X_0}(\mathbf{x}_i)$	$\underline{p_5}$ $D^2_{X_1}(\mathbf{x}_i)$
1	$D_{X_0}(\mathbf{x}_1)$	$D_{X_1}(\mathbf{x}_1)$	$D_{X_0}(\mathbf{x}_1) \cdot D_{X_1}(\mathbf{x}_1)$	$D^2_{X_0}(\mathbf{x}_1)$	$D^2_{X_1}(\mathbf{x}_1)$
2	$D_{X_0}(\mathbf{x}_2)$	$D_{X_1}(\mathbf{x}_2)$	$D_{X_0}(\mathbf{x}_2) \cdot D_{X_1}(\mathbf{x}_2)$	$D^2_{X_0}(\mathbf{x}_2)$	$D^2_{X_1}(\mathbf{x}_2)$
...					
i	$D_{X_0}(\mathbf{x}_i)$	$D_{X_1}(\mathbf{x}_i)$	$D_{X_0}(\mathbf{x}_i) \cdot D_{X_1}(\mathbf{x}_i)$	$D^2_{X_0}(\mathbf{x}_i)$	$D^2_{X_1}(\mathbf{x}_i)$
...					
$m+n$	$D_{X_0}(\mathbf{x}_{m+n})$	$D_{X_1}(\mathbf{x}_{m+n})$	$D_{X_0}(\mathbf{x}_{m+n}) \cdot D_{X_1}(\mathbf{x}_{m+n})$	$D^2_{X_0}(\mathbf{x}_{m+n})$	$D^2_{X_1}(\mathbf{x}_{m+n})$

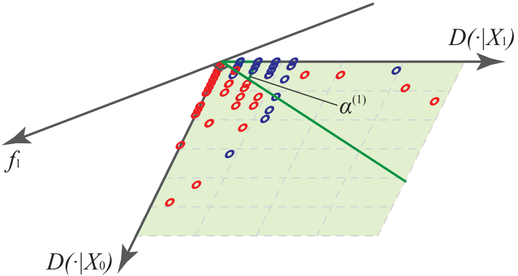
$DD\alpha$-classifier



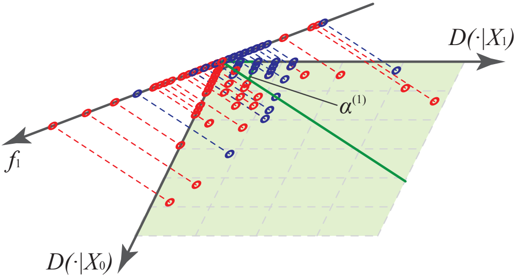
$DD\alpha$-classifier



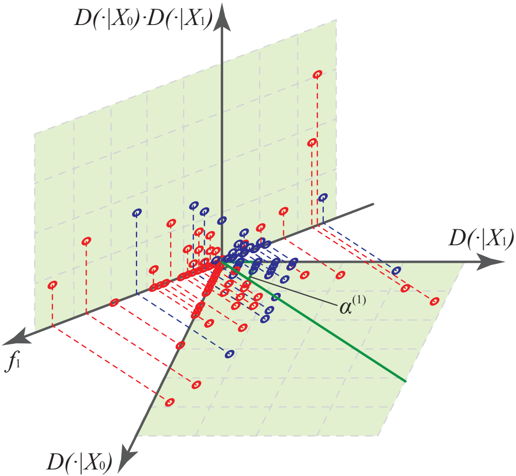
$DD\alpha$-classifier



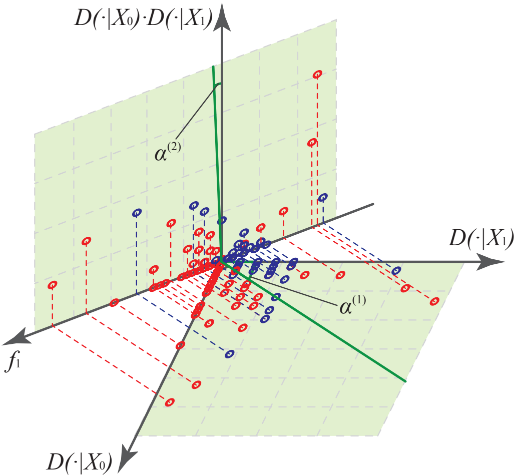
$DD\alpha$-classifier



$DD\alpha$-classifier



$DD\alpha$-classifier



Contents

Depth-based classification

Data depth $+$ **Classification**

$=$

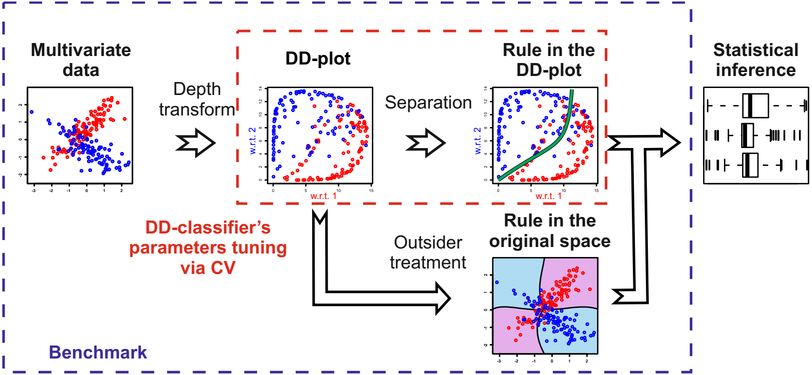
affine-invariante robust non-parametric distribution-free
classification

Problems:

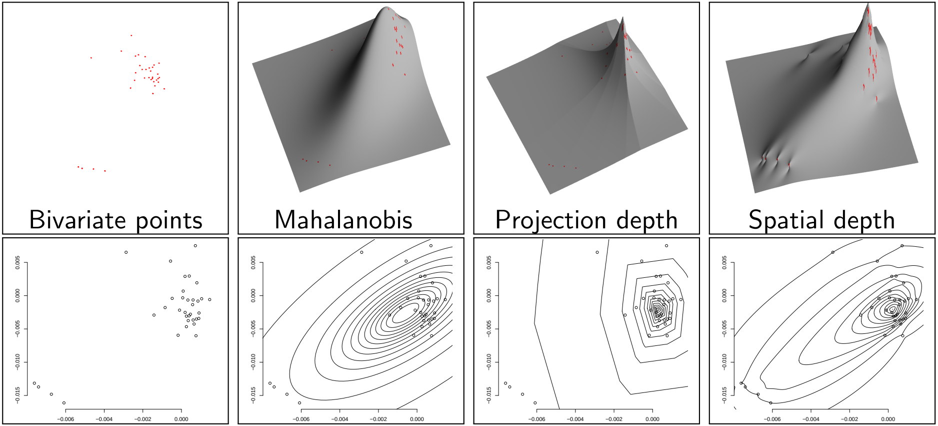
- lack of implementations;
- different languages and interfaces;
- different requirements to the format of the input data;
- no implementations of depths and *DD*-classifiers under one roof.

We summarize the work of many researchers.

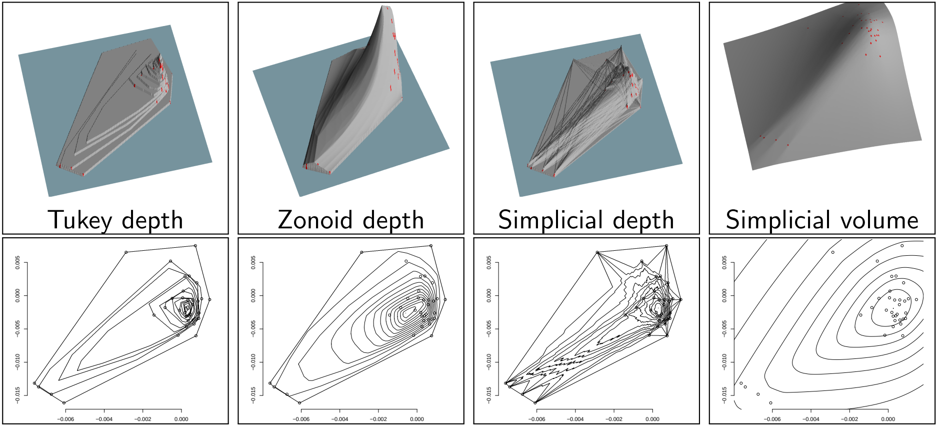
R-package ddalpha is a structured solution



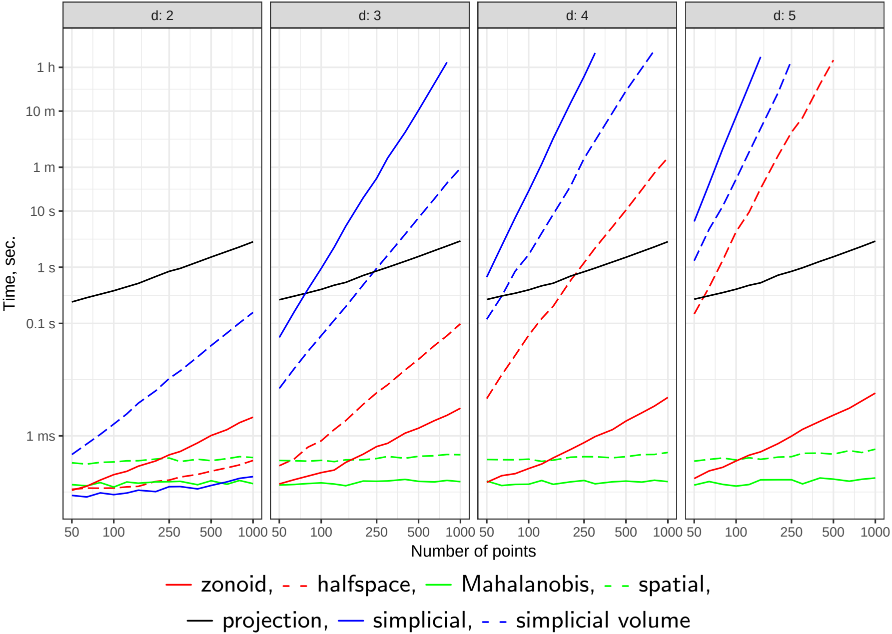
Implemented data depths



Implemented data depths



Implemented data depths: computation time



Implemented data depths: algorithms

Depth	Exact	Approximate
Mahalanobis	✓	✓robust(mcd)
projection		✓pp + ✓Nelder-Mead
spatial (L_1)	✓	✓robust(mcd)
halfspace	✓✓✓	✓pp
zonoid	✓	
simplicial	✓	✓part of simplices
simplicial volume	✓	✓part of simplices

Contents

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 Rcpp (>= 0.11.0)

LinkingTo BH, Rcpp

Description Contains procedures for depth-based supervised learning, which are entirely non-parametric, in particular the DDalpha-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?

- Pokotylo, O., Mozharovskyi, P., Dyckerhoff, R. (2017).
  Depth and depth-based classification with R-package ddalpha.
  Journal of Statistical Software, in press.
- Nagy, S., Gijbels, I., Hlubinka, D. (2017).
  Depth-based recognition of shape outlying functions.
  Journal of Computational and Graphical Statistics, 26, 883–893.

- Dyckerhoff R., Mosler K., Koshevoy G. (1996).
  Zonoid data depth: Theory and computation.
  In A Prat (ed.), *COMPSTAT '96 – Proceedings in Computational Statistics*, pp. 235–240. Springer.
- Lange T., Mosler K., Mozharovskyi P. (2014).
  Fast nonparametric classification based on data depth.
  Statistical Papers, 55, 49–69.
- Dyckerhoff R., Mozharovskyi P. (2016).
  Exact computation of the halfspace depth.
  Computational Statistics and Data Analysis, 98, 19–30.
- Pokotylo O., Mosler K. (2016).
  Classification with the pot-pot plot.
  Statistical Papers, to appear.