DD-Plots

Christine Müller

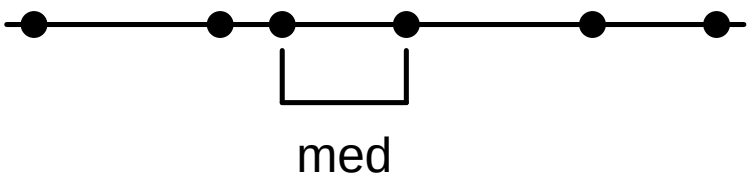
Fakultät Statistik

Technische Universität Dortmund

technische universität dortmund

What is data depth?

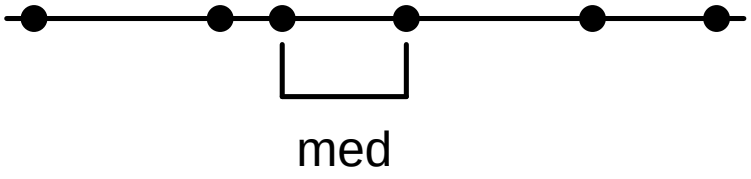
Median:



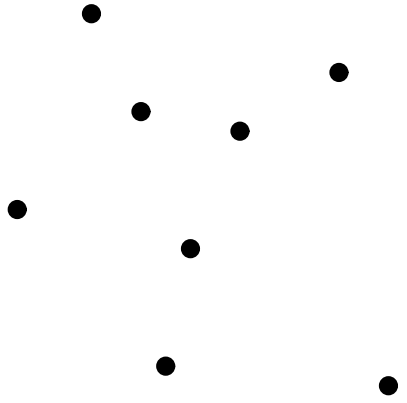
1. Data depth as generalization of the median

What is data depth?

Median:

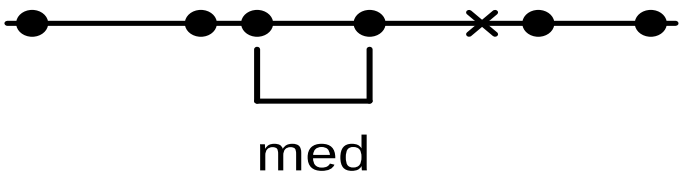


Generalization for multivariate data?



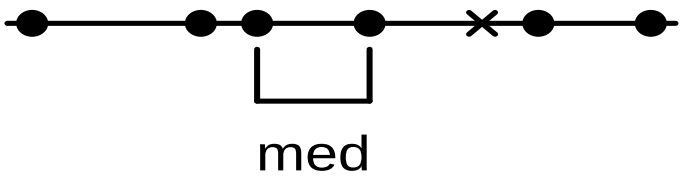
Data depth of μ in $z_* = (z_1, \ldots, z_N)$ (univariate):

$$d(\mu, z_*) = \frac{1}{N} \min \{\sharp\{n;\ z_n \leq \mu\}, \sharp\{n;\ z_n \geq \mu\}\}$$



Data depth of μ in $z_* = (z_1, \ldots, z_N)$ (univariate):

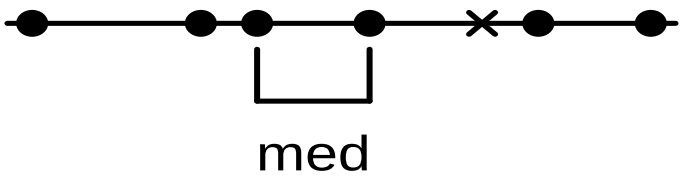
$$d(\mu, z_*) = \frac{1}{N} \min \{\sharp\{n;\ z_n \leq \mu\}, \sharp\{n;\ z_n \geq \mu\}\}$$



Here $d(\mu, z_*) = \frac{2}{6}$

Data depth of μ in $z_* = (z_1, \ldots, z_N)$ (univariate):

$$d(\mu, z_*) = \frac{1}{N} \min \{\sharp\{n;\ z_n \leq \mu\}, \sharp\{n;\ z_n \geq \mu\}\}$$

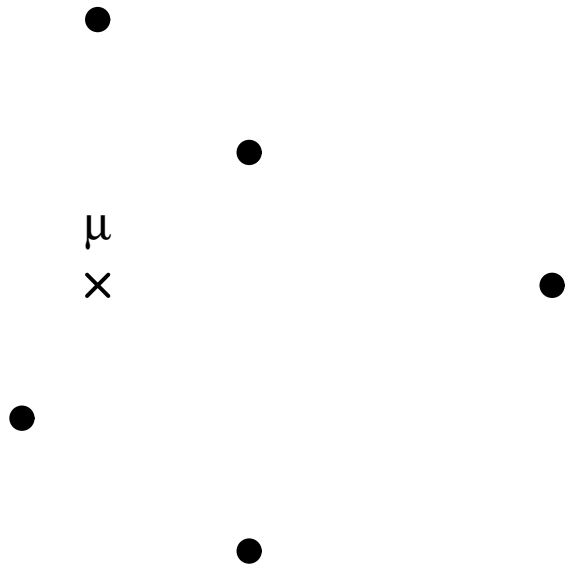


Here $d(\mu, z_*) = \frac{2}{6}$

Median: all μ with maximum data depth, i.e. all μ with $d(\mu, z_*) = \frac{1}{2}$

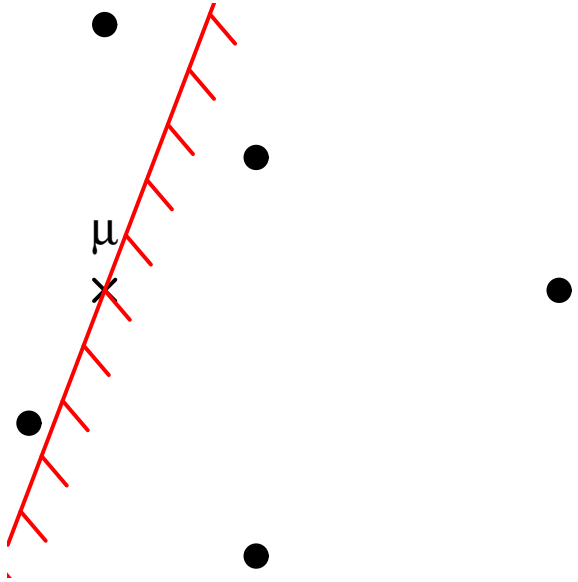
Data depth of μ in $z_* = (z_1, \ldots, z_N)$ (multivariate):
Half space depth of Tukey 1975

$$d_H(\mu, z_*) = \frac{1}{N} \min_H \sharp\{n;\ z_n \text{ lies in a half space } H \text{ containing } \mu\}$$



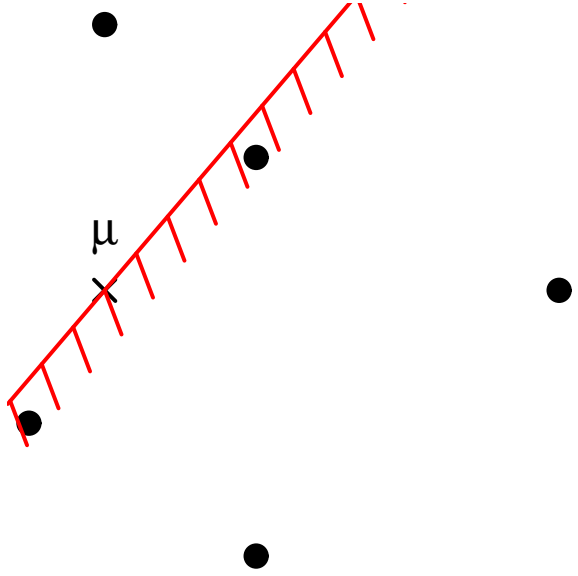
Data depth of μ in $z_* = (z_1, \ldots, z_N)$ (multivariate):
Half space depth of Tukey 1975

$$d_H(\mu, z_*) = \frac{1}{N} \min_H \sharp\{n;\ z_n \text{ lies in a half space } H \text{ containing } \mu\}$$



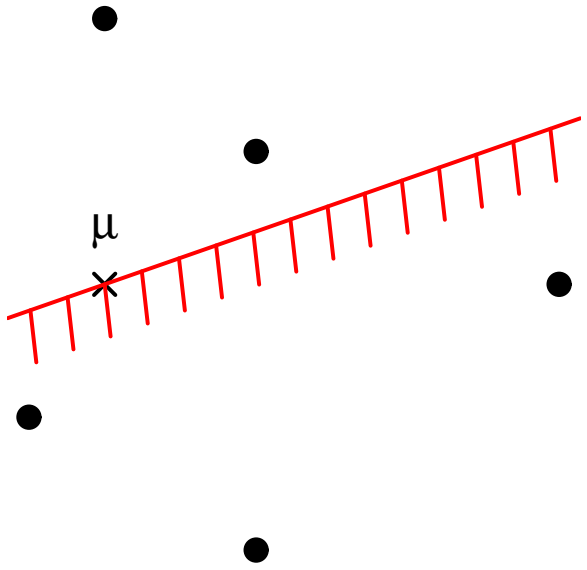
Data depth of μ in $z_* = (z_1, \ldots, z_N)$ (multivariate):
Half space depth of Tukey 1975

$$d_H(\mu, z_*) = \frac{1}{N} \min_H \sharp\{n;\ z_n \text{ lies in a half space } H \text{ containing } \mu\}$$



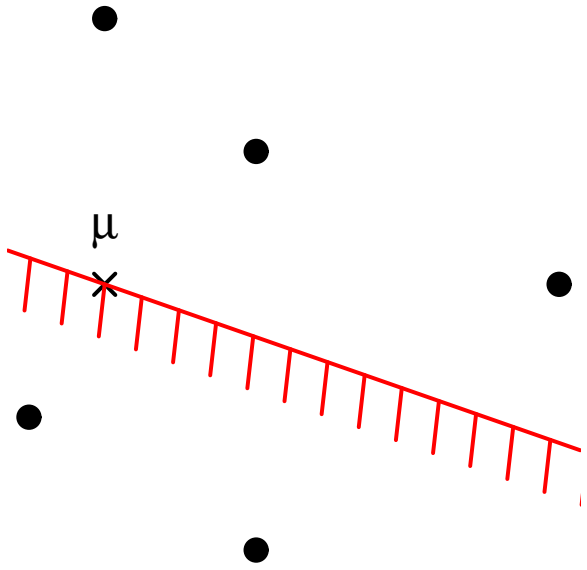
Data depth of μ in $z_* = (z_1, \ldots, z_N)$ (multivariate):
Half space depth of Tukey 1975

$$d_H(\mu, z_*) = \frac{1}{N} \min_H \sharp\{n;\ z_n \text{ lies in a half space } H \text{ containing } \mu\}$$



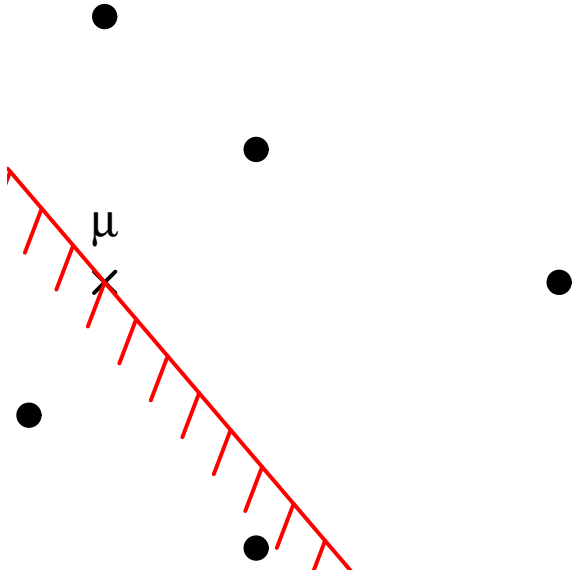
Data depth of μ in $z_* = (z_1, \ldots, z_N)$ (multivariate):
Half space depth of Tukey 1975

$$d_H(\mu, z_*) = \frac{1}{N} \min_H \sharp\{n;\ z_n \text{ lies in a half space } H \text{ containing } \mu\}$$



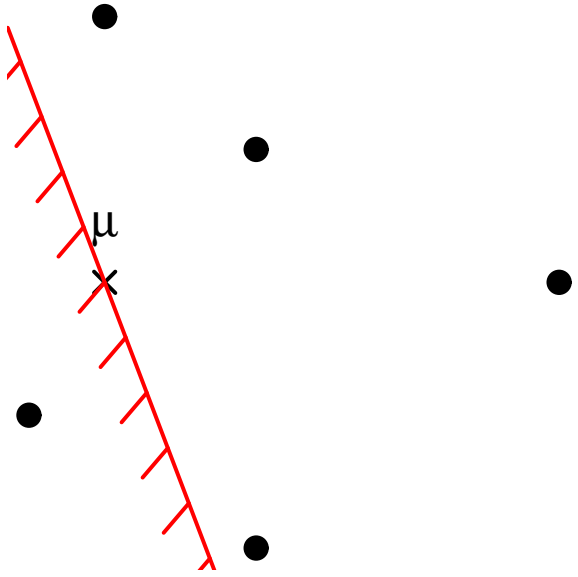
Data depth of μ in $z_* = (z_1, \ldots, z_N)$ (multivariate):
Half space depth of Tukey 1975

$$d_H(\mu, z_*) = \frac{1}{N} \min_H \sharp\{n;\ z_n \text{ lies in a half space } H \text{ containing } \mu\}$$



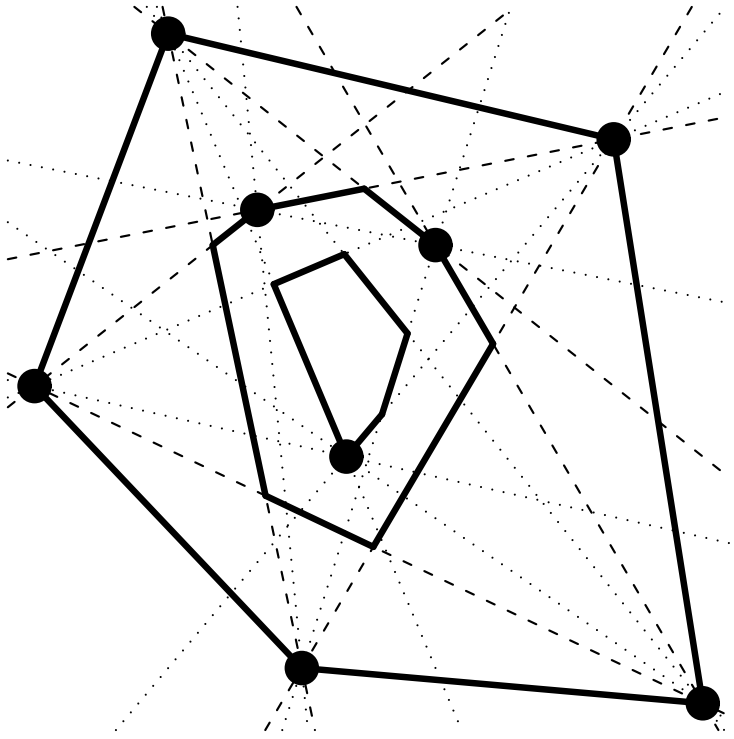
Data depth of μ in $z_* = (z_1, \ldots, z_N)$ (multivariate):
Half space depth of Tukey 1975

$$d_H(\mu, z_*) = \frac{1}{N} \min_H \sharp\{n;\ z_n \text{ lies in a half space } H \text{ containing } \mu\}$$

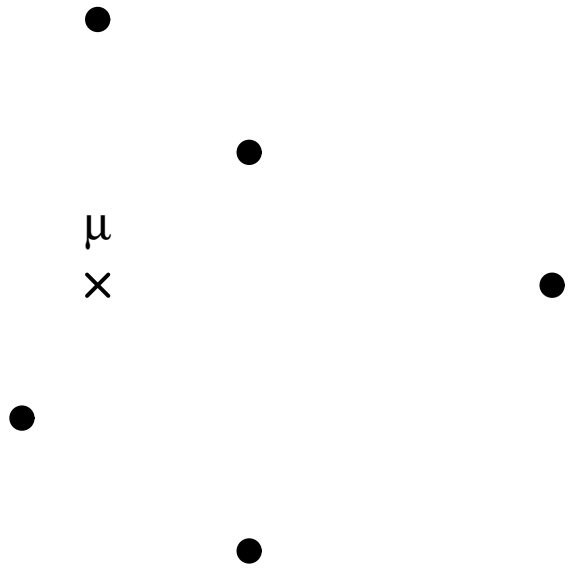


Here: $d_H(\mu, z_*) = \frac{1}{5} \cdot 1$

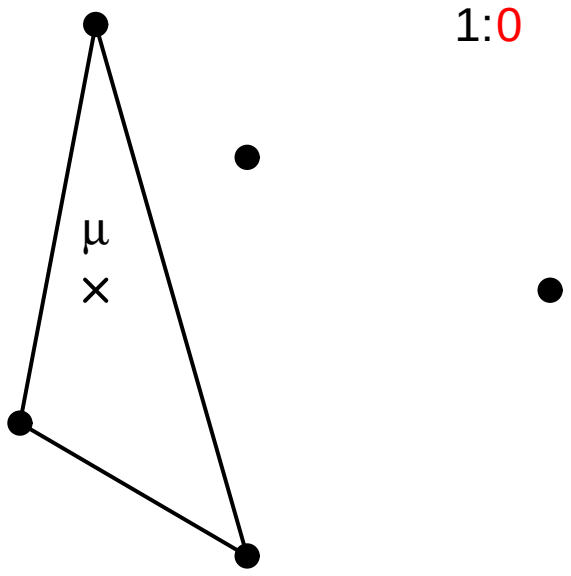
Half space median / Tukey median:
all μ with maximum half space depth



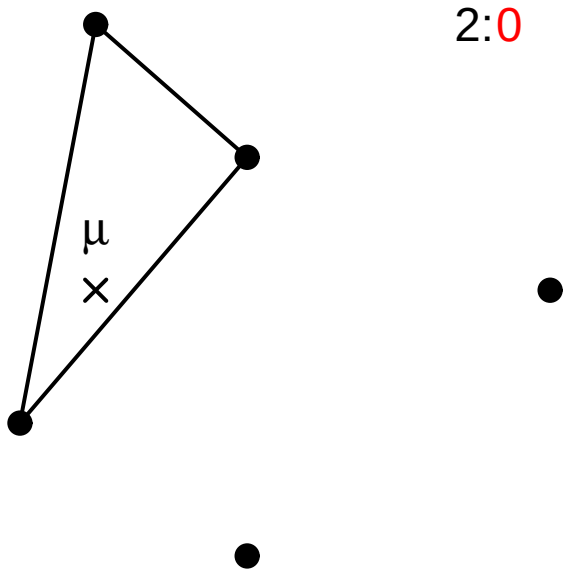
Simplicial depth $d_S(\mu, z_*)$ of a location parameter μ (Liu 1988,1990):
The relative number of simplicies (triangles) containing μ



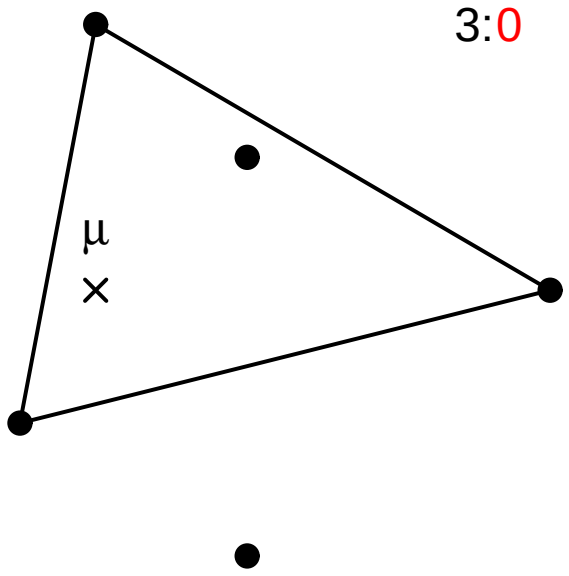
Simplicial depth $d_S(\mu, z_*)$ of a location parameter μ (Liu 1988,1990):
The relative number of simplicies (triangles) containing μ



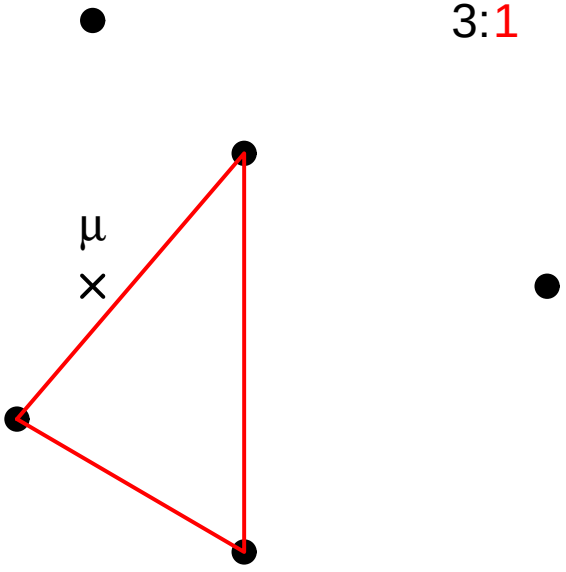
Simplicial depth $d_S(\mu, z_*)$ of a location parameter μ (Liu 1988,1990):
The relative number of simplicies (triangles) containing μ



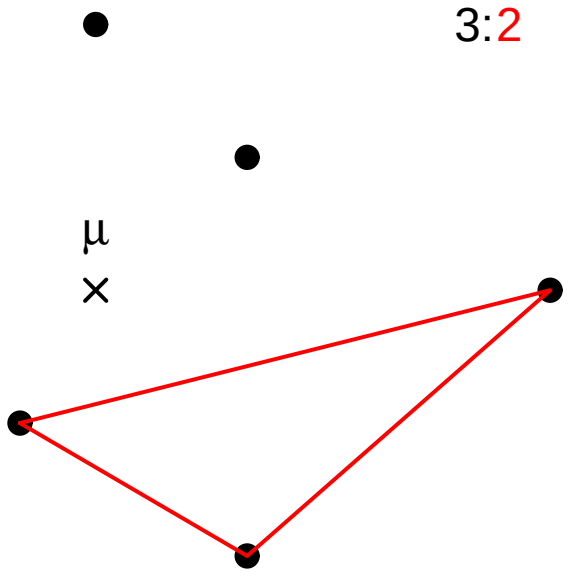
Simplicial depth $d_S(\mu, z_*)$ of a location parameter μ (Liu 1988,1990):
The relative number of simplicies (triangles) containing μ



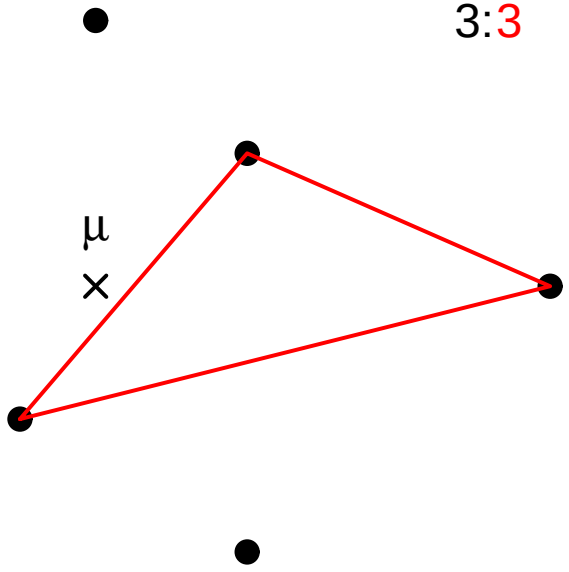
Simplicial depth $d_S(\mu, z_*)$ of a location parameter μ (Liu 1988,1990):
The relative number of simplicies (triangles) containing μ



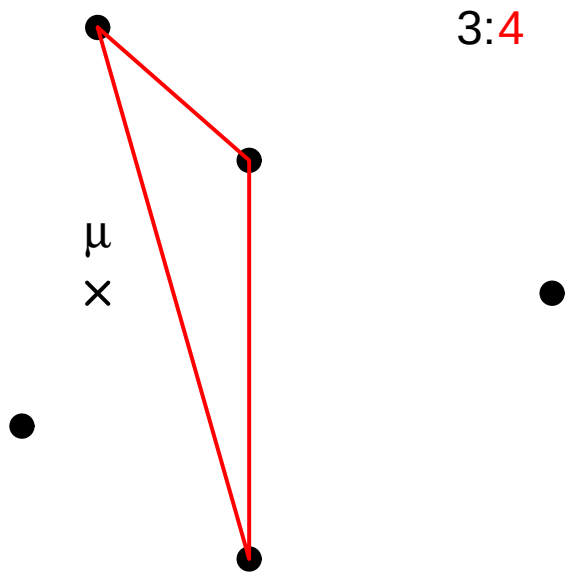
Simplicial depth $d_S(\mu, z_*)$ of a location parameter μ (Liu 1988,1990):
The relative number of simplicies (triangles) containing μ



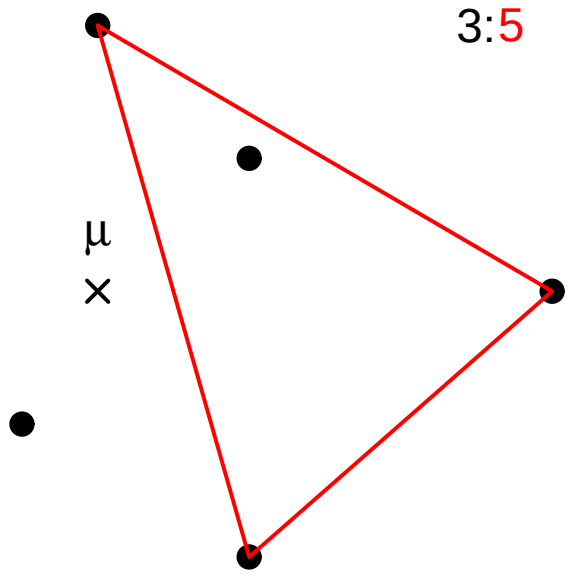
Simplicial depth $d_S(\mu, z_*)$ of a location parameter μ (Liu 1988,1990):
The relative number of simplicies (triangles) containing μ



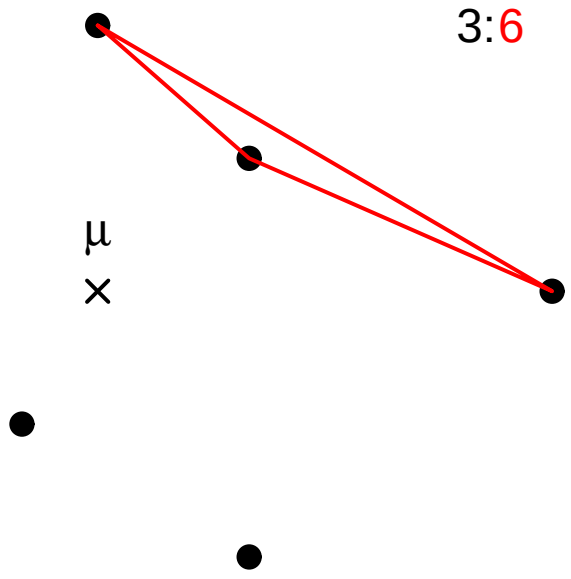
Simplicial depth $d_S(\mu, z_*)$ of a location parameter μ (Liu 1988,1990):
The relative number of simplicies (triangles) containing μ



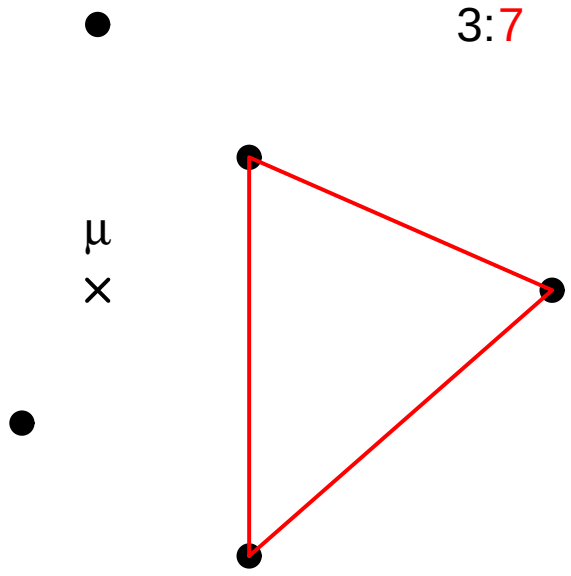
Simplicial depth $d_S(\mu, z_*)$ of a location parameter μ (Liu 1988,1990):
The relative number of simplicies (triangles) containing μ



Simplicial depth $d_S(\mu, z_*)$ of a location parameter μ (Liu 1988,1990):
The relative number of simplicies (triangles) containing μ



Simplicial depth $d_S(\mu, z_*)$ of a location parameter μ (Liu 1988,1990):
The relative number of simplicies (triangles) containing μ



Here $d_S(\mu, z_*) = \frac{1}{\binom{5}{3}} \cdot 3 = \frac{1}{10} \cdot 3$

$= \frac{1}{\binom{N}{3}} \cdot \sharp\{\{n_1, n_2, n_3\};\ \mu \in$ simplex spanned by $z_{n_1}, z_{n_2}, z_{n_3}\}$

Group 0: $(z_1, \ldots, z_N)$

Group 1: $(z_{N+1}, \ldots, z_{N+M})$

$z_n \in I\!R^K$ for $n = 1, \ldots, N + M$

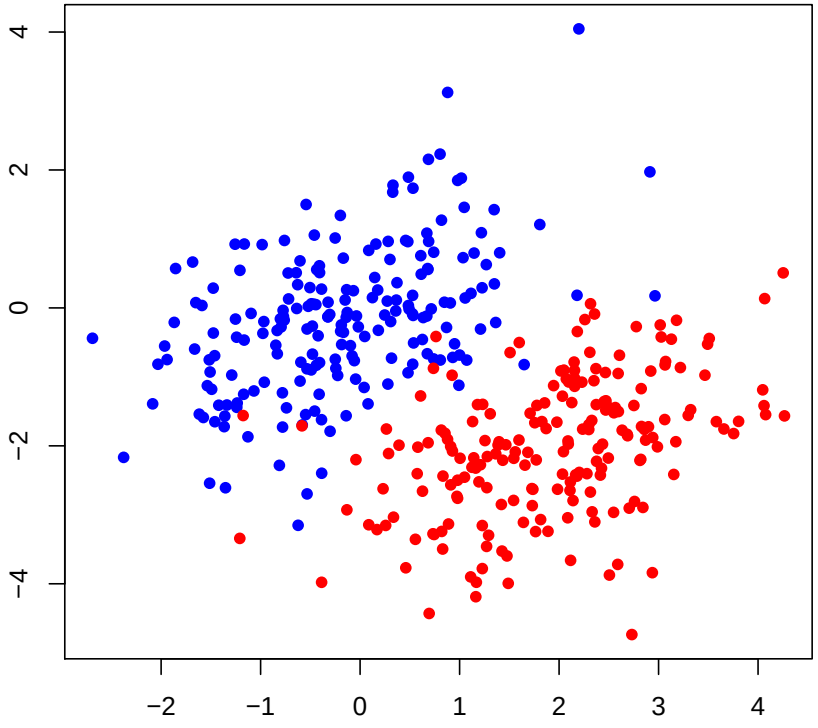
Depth-Depth-Plot (Li et al. 2012):

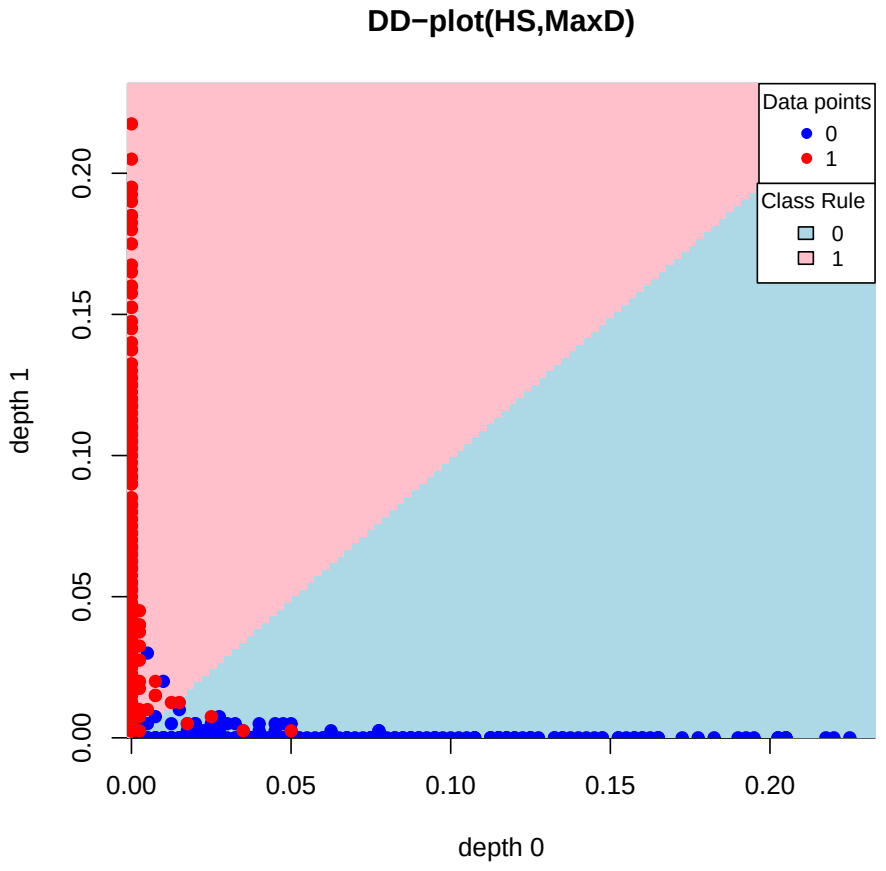
Plot of $(d(z_n, (z_1, \ldots, z_N)),\ d(z_n, (z_{N+1}, \ldots, z_{N+M})))$

for $n = 1, \ldots, N + M$

technische universität dortmund

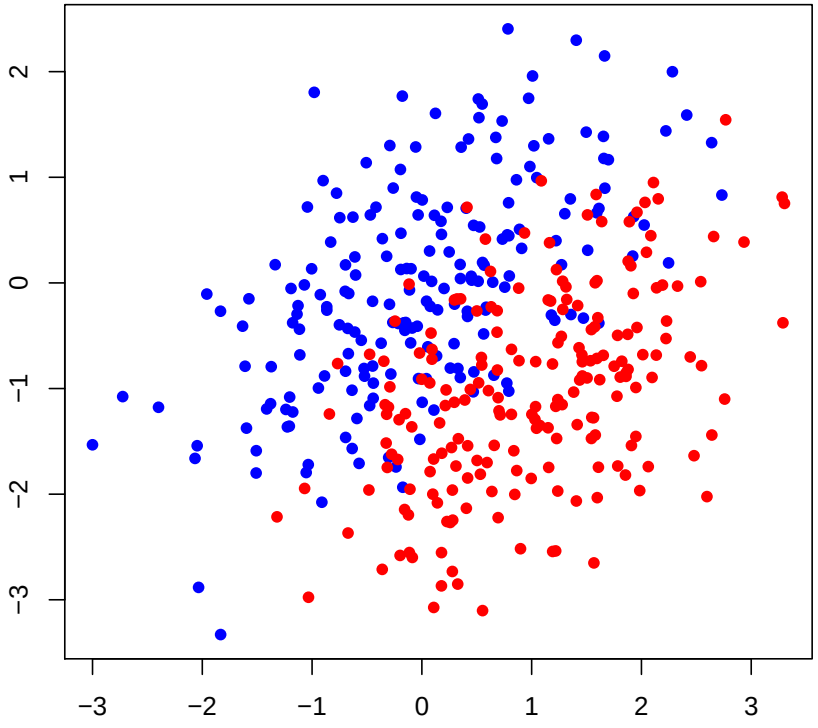
Scatter-Plot and Depth-Depth-Plot for $K = 2$: Almost separated groups

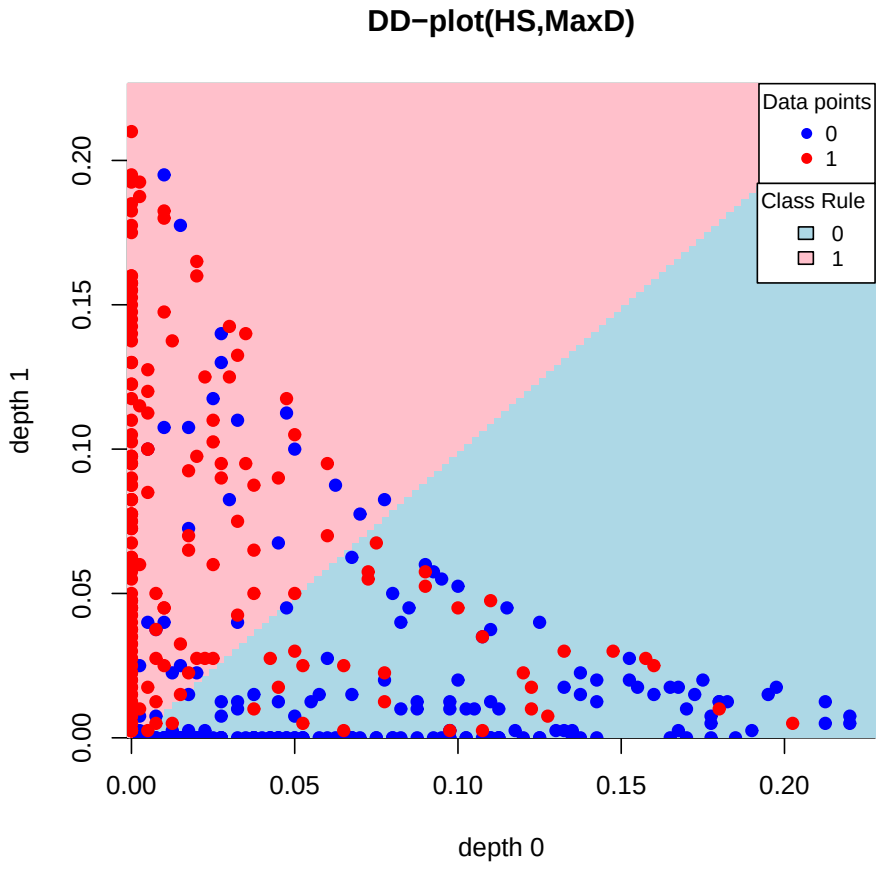




technische universität dortmund

Scatter-Plot and Depth-Depth-Plot for $K = 2$: Overlapping groups





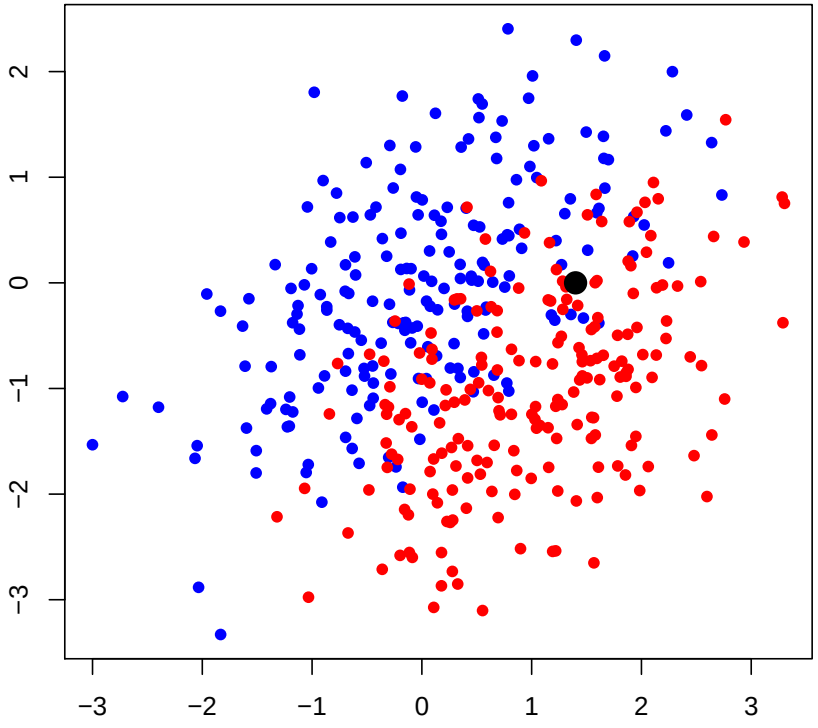
Classifying a new observation z_0:

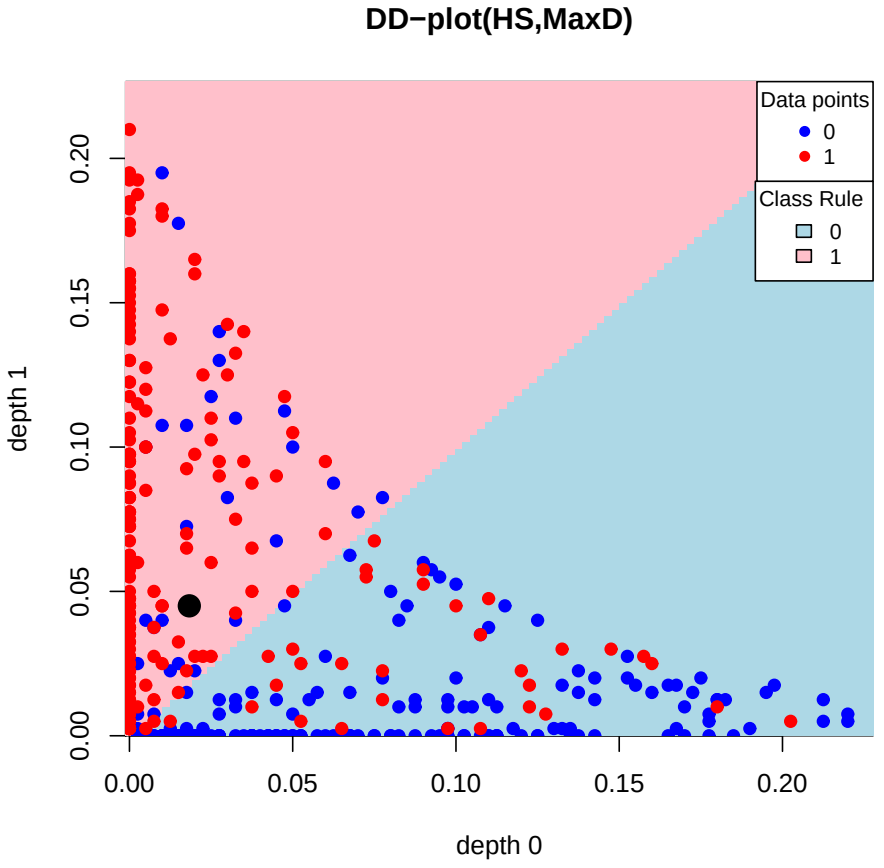
$$d(z_0, (z_1, \ldots, z_N)) > d(z_0, (z_{N+1}, \ldots, z_{N+M})) \longrightarrow \text{Group 0}$$

$$d(z_0, (z_1, \ldots, z_N)) < d(z_0, (z_{N+1}, \ldots, z_{N+M})) \longrightarrow \text{Group 1}$$

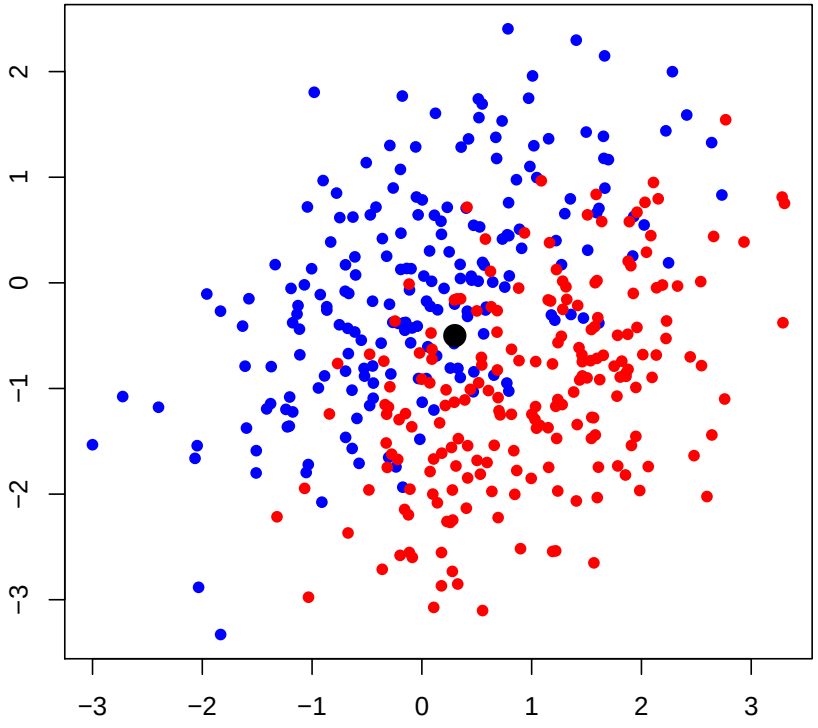
technische universität dortmund

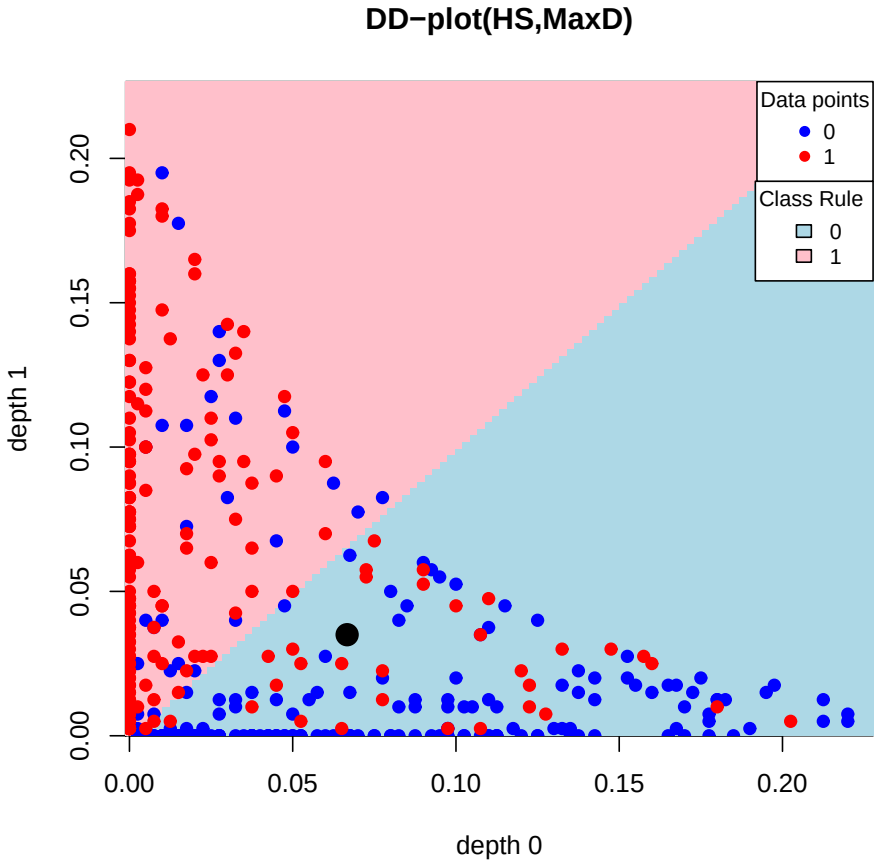
Classifying a new observation: new observation $\longrightarrow$ group 1



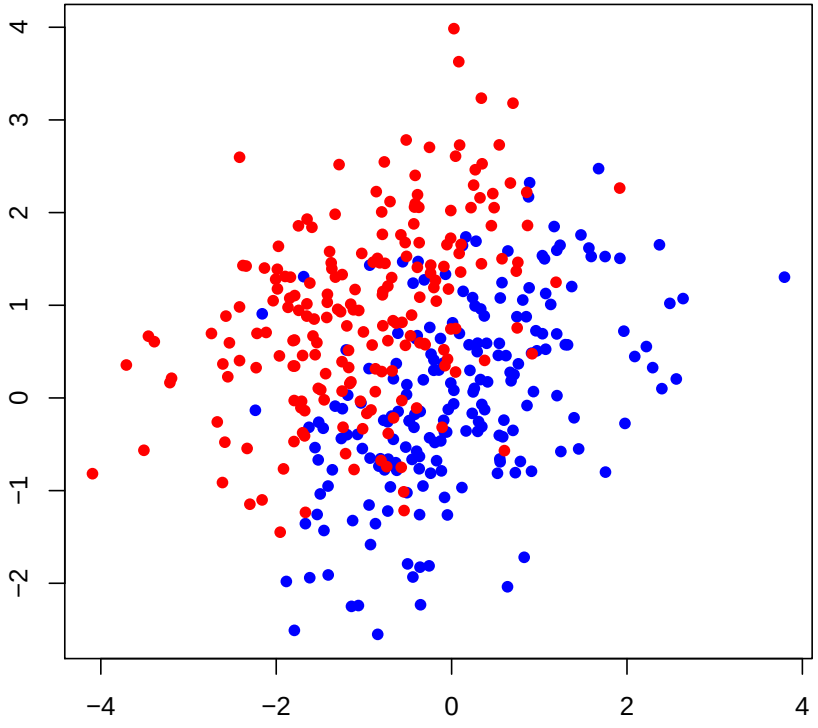


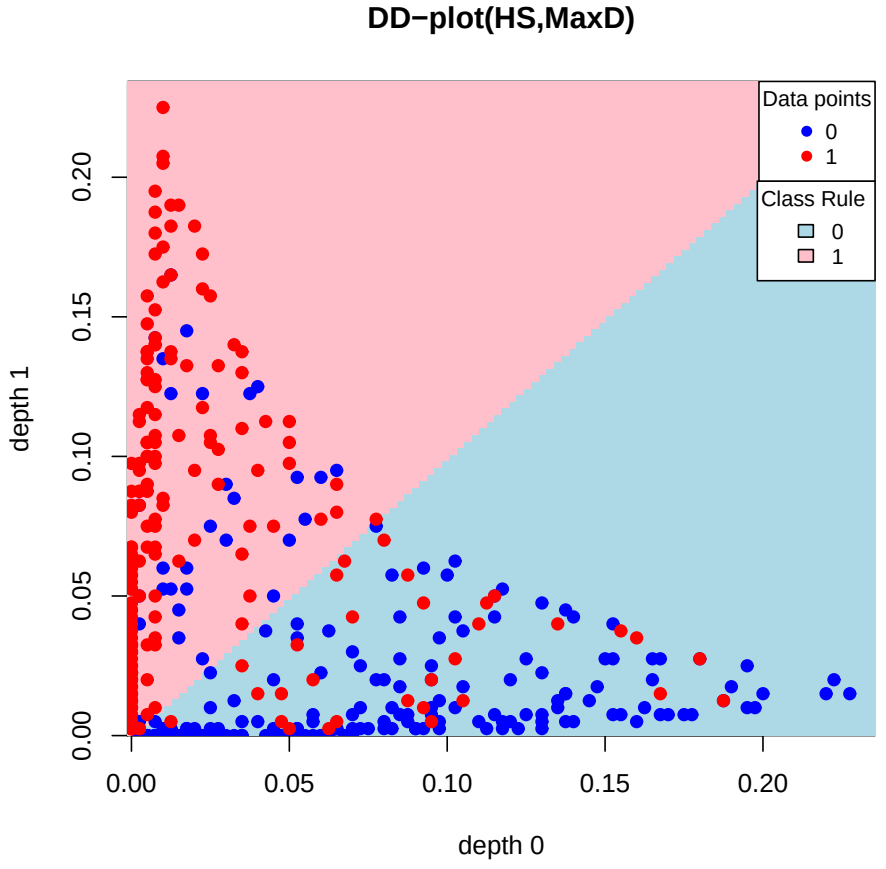
Classifying a new observation: new observation $\longrightarrow$ group 0



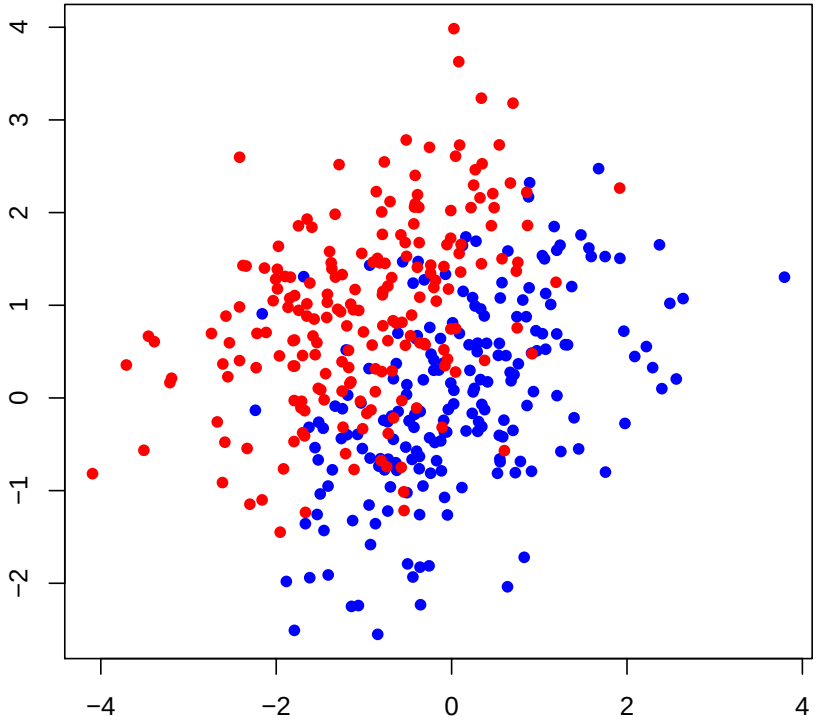


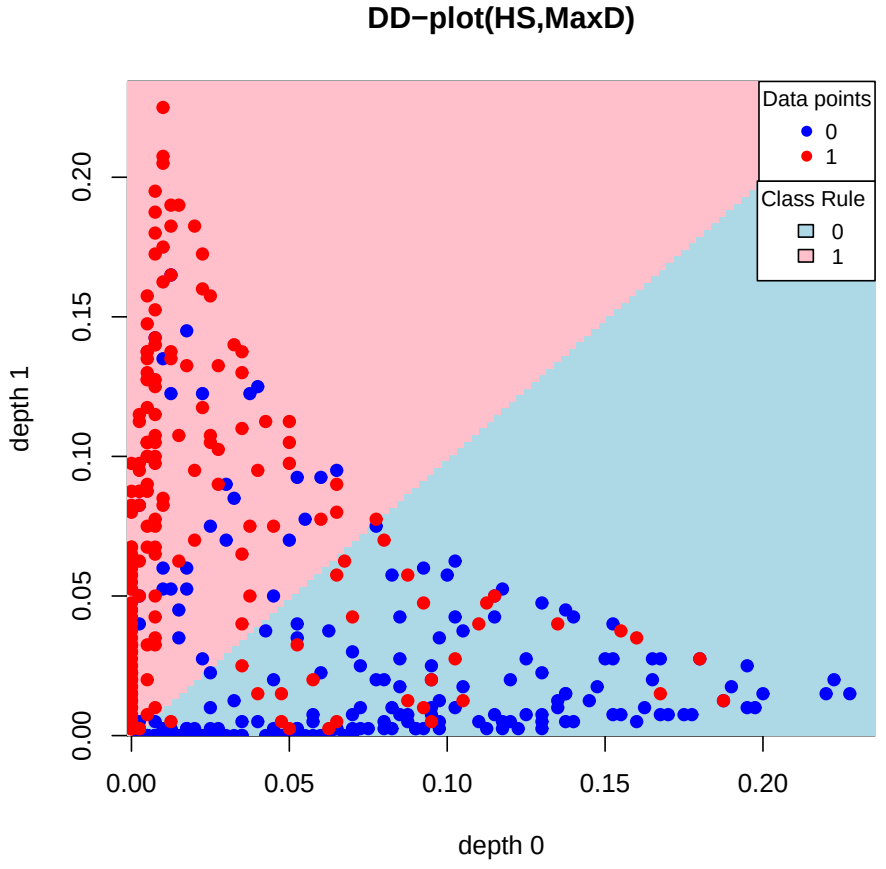
Prediction error: DD-HS=0.186



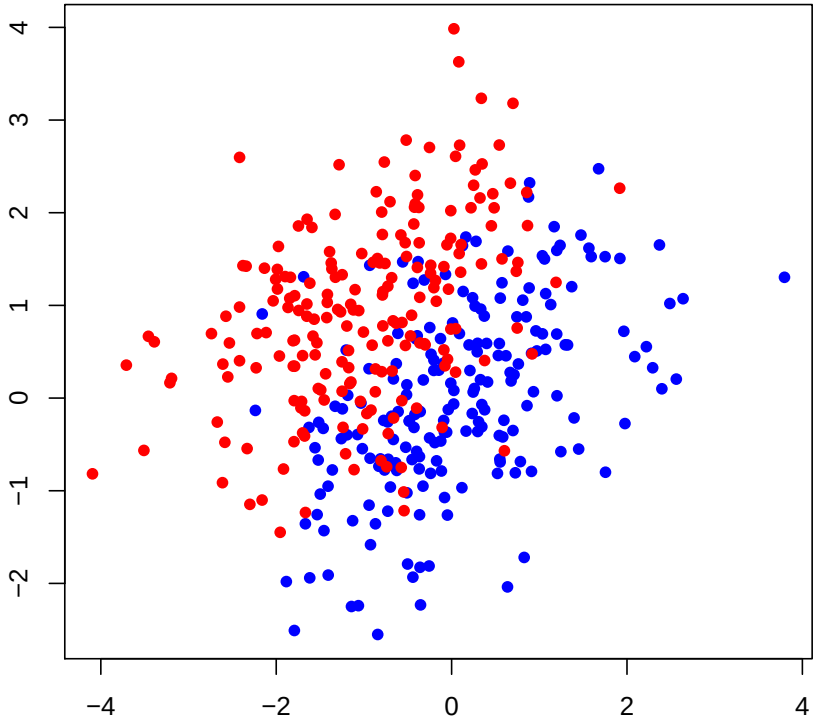


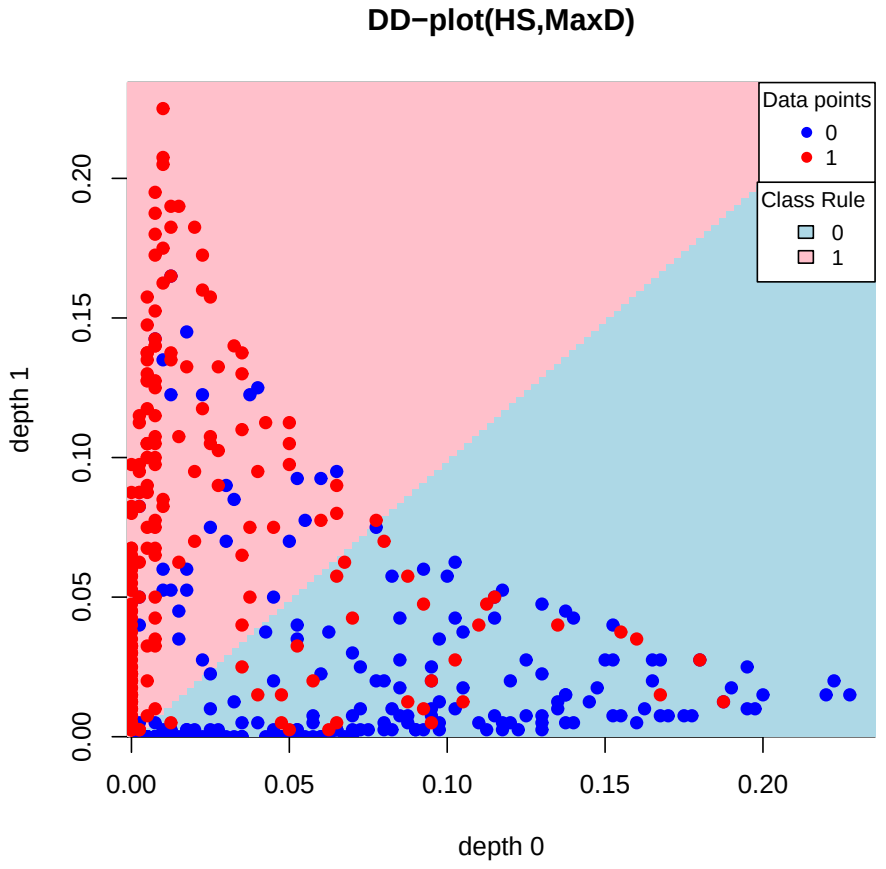
Prediction error: DD-HS=0.186, DD-SD=0.179





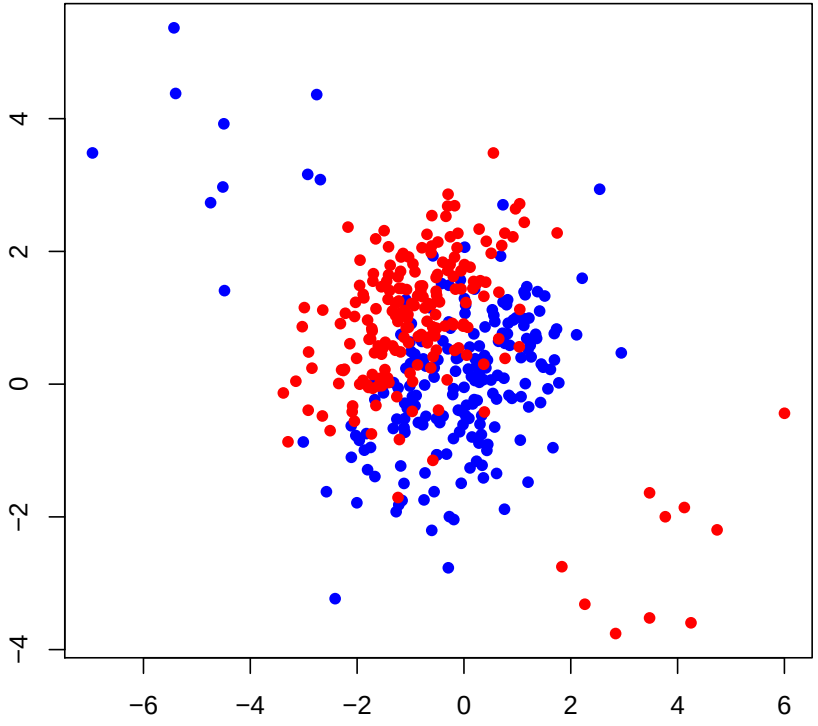
Prediction error: DD-HS=0.186, DD-SD=0.179, LDA=0.173, QDA=0.175, RF=0.21

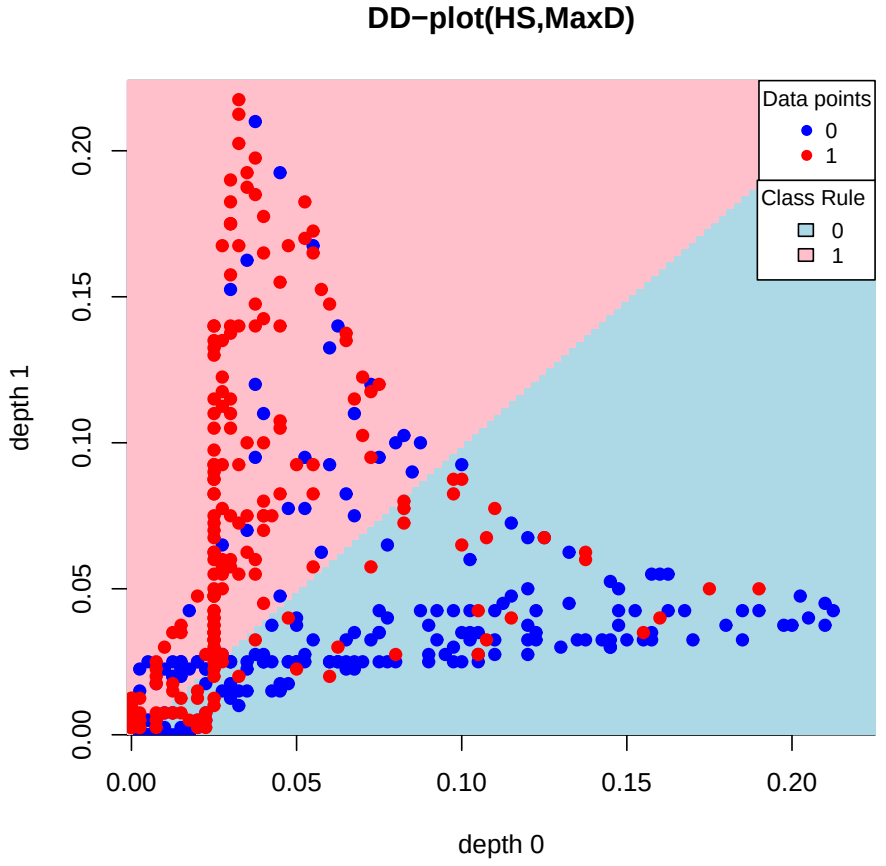




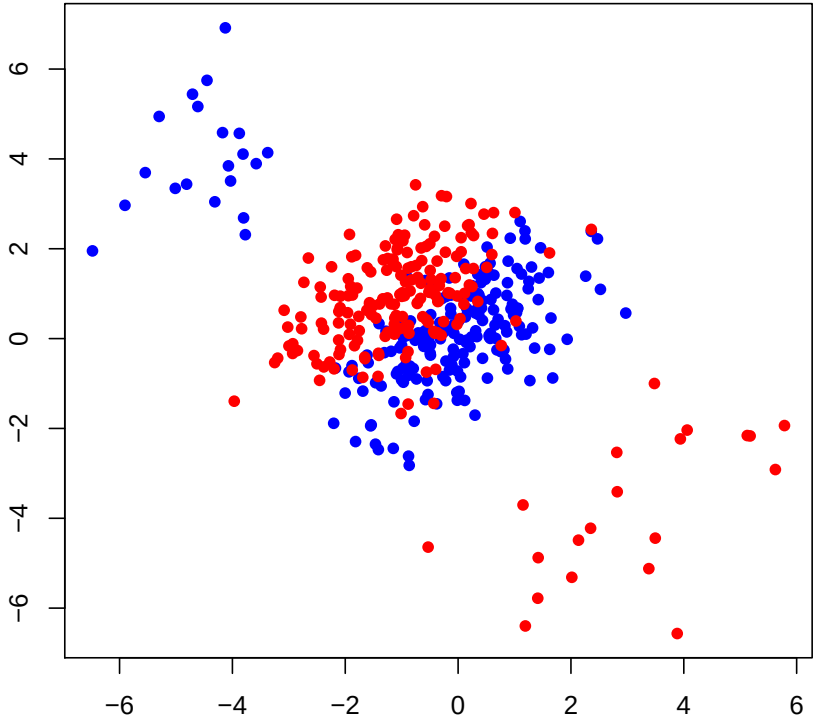
technische universität dortmund

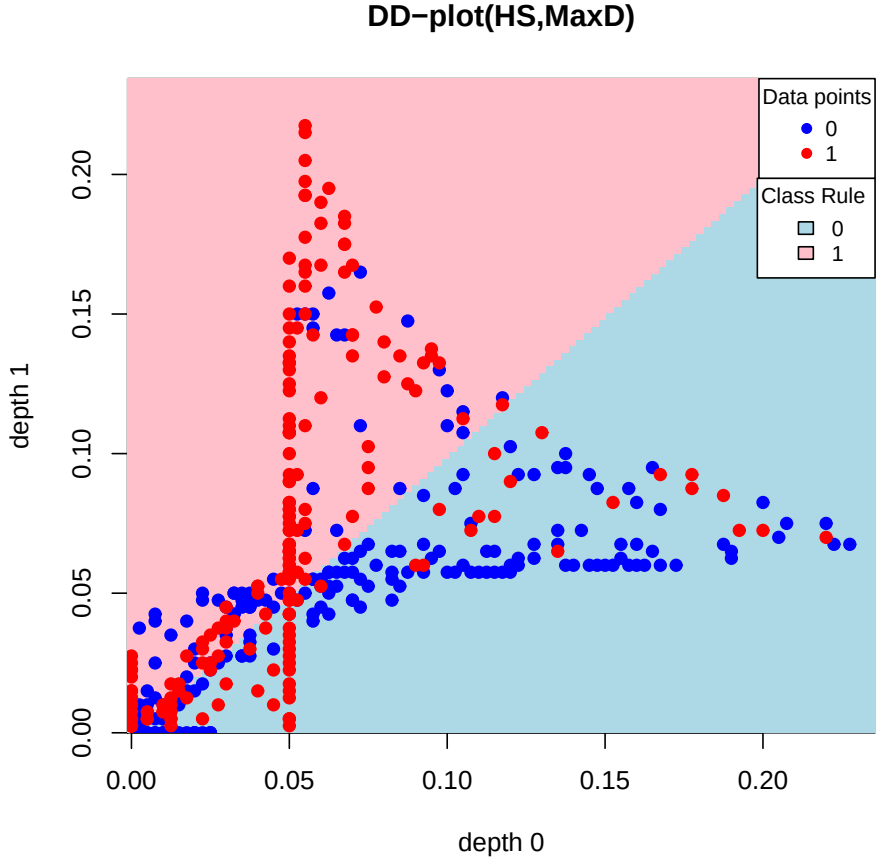
Prediction error (5% contamination):
DD-HS=0.306, DD-SD=0.297,
LDA=0.218, QDA=0.238, RF=0.17



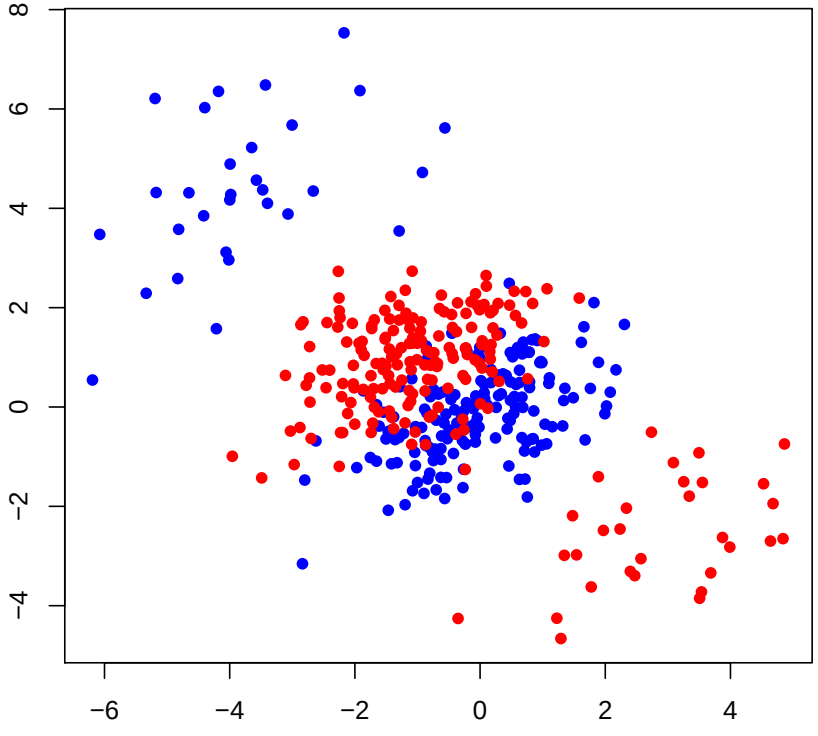


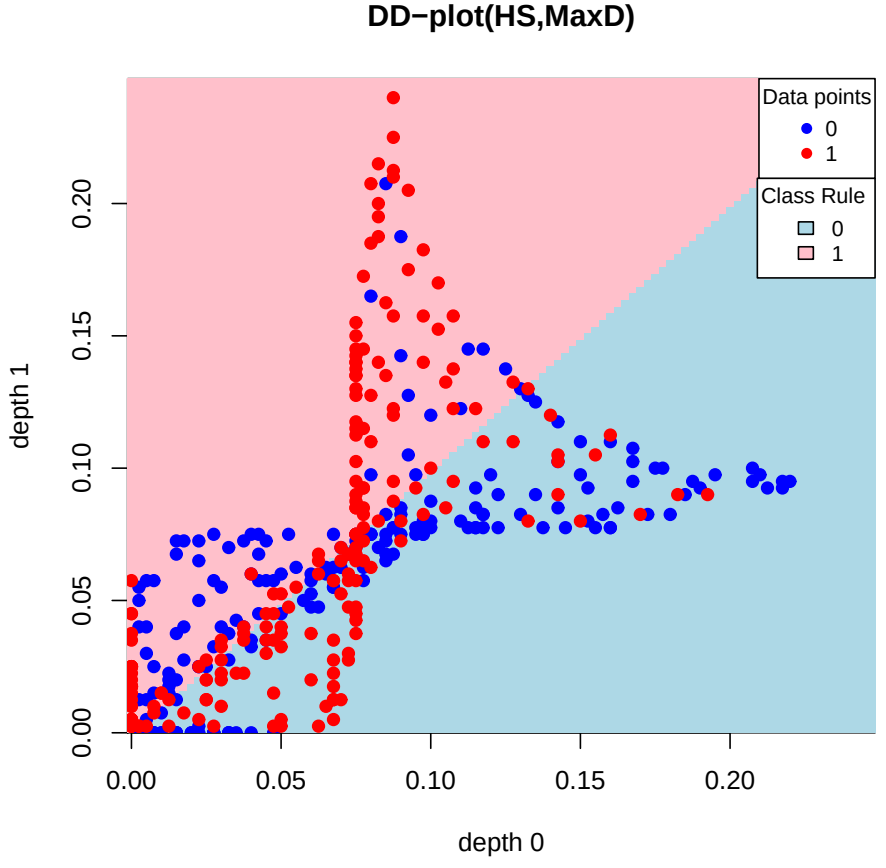
Prediction error (10% contamination):
DD-HS=0.4, DD-SD=0.35,
LDA=0.439, QDA=0.466, RF=0.176





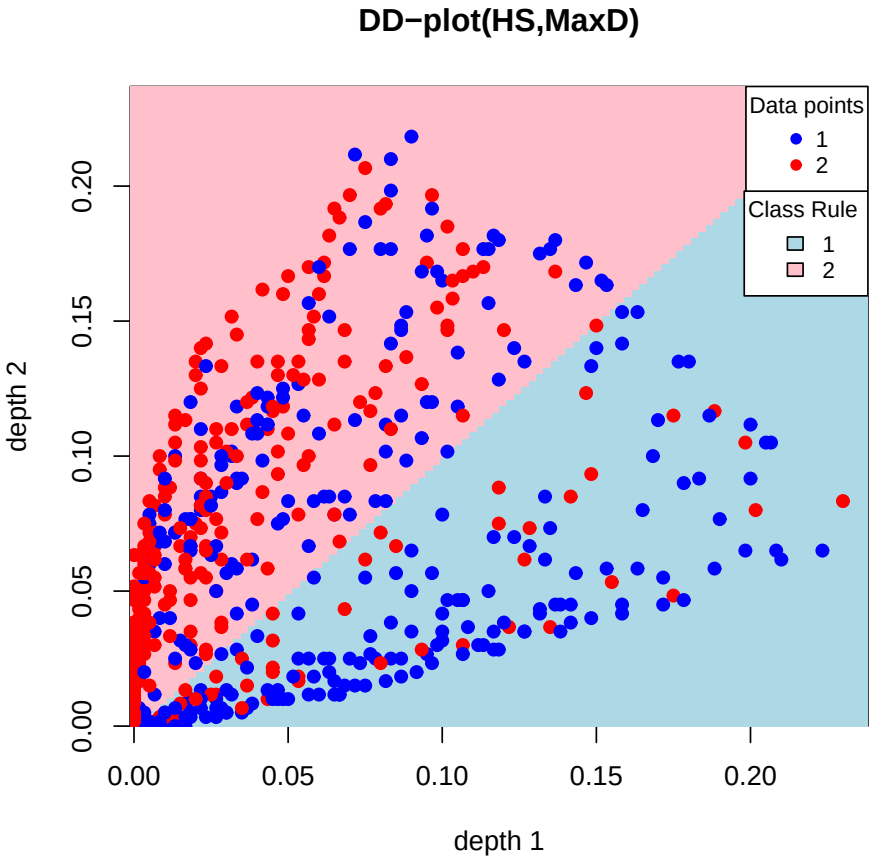
Prediction error (15% contamination):
DD-HS=0.511, DD-SD=0.424,
LDA=0.836, QDA=0.554, RF=0.184

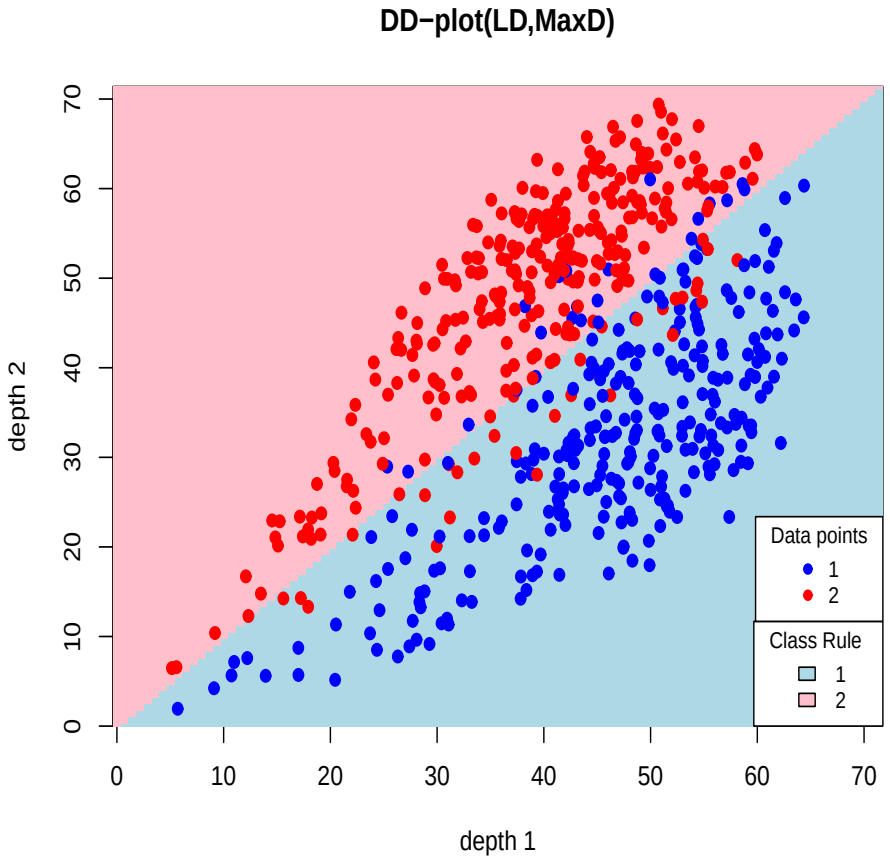




DD-plot for structure-borne noise data at two time points using maxDepth with

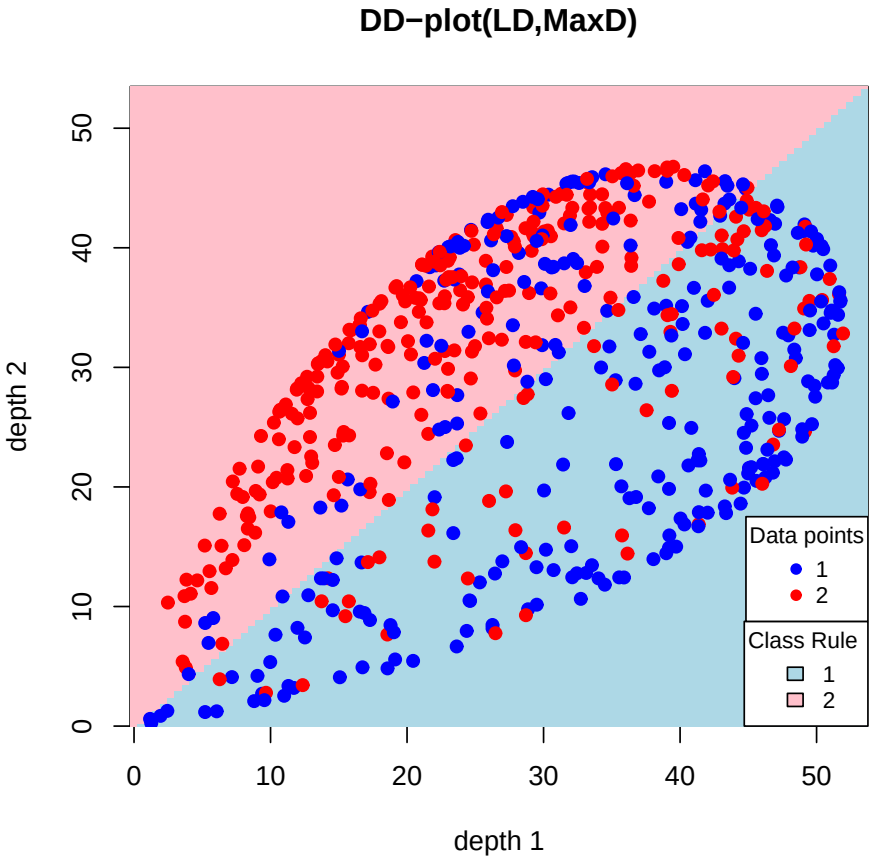
2 principal components or 6 principal components

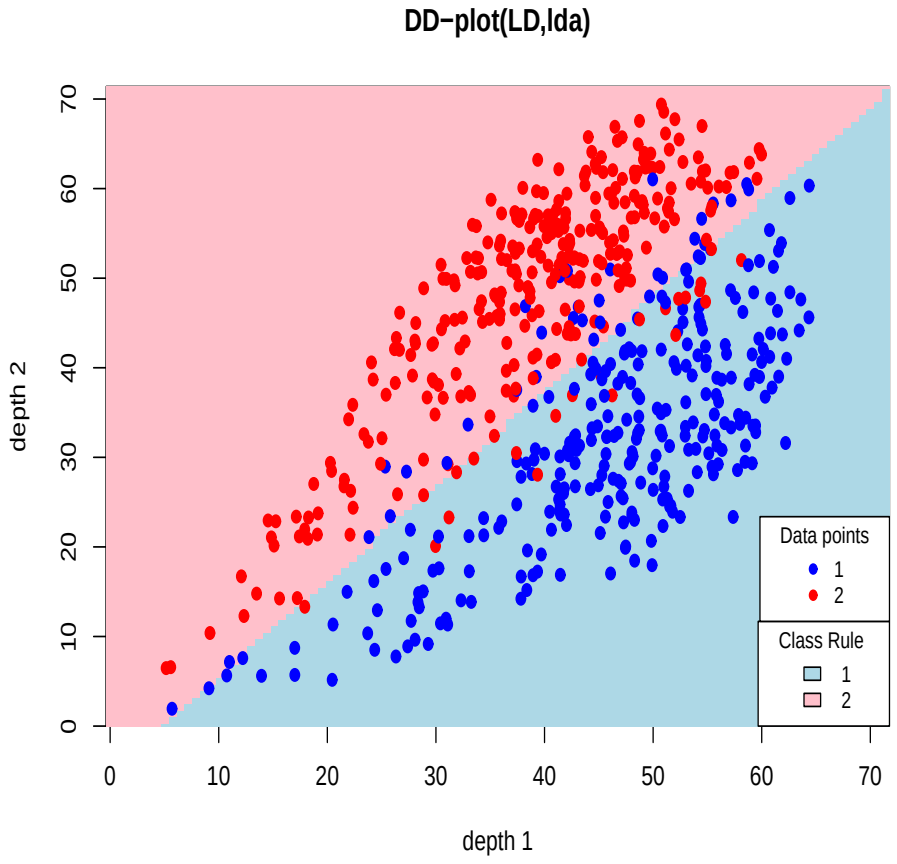




DD-plot for structure-borne noise data at two time points using LDA with

2 principal components or 6 principal components





- DD-classification can be better than LDA, QDA, and RF.

- DD-classification can be better than LDA, QDA, and RF.
- DD-classification based on simplicial depth is slightly better than that based on half-space depth.

3. Data depth for classification

- DD-classification can be better than LDA, QDA, and RF.
- DD-classification based on simplicial depth is slightly better than that based on half-space depth.
- DD-classification can be based on other depth notions.

- DD-classification can be better than LDA, QDA, and RF.
- DD-classification based on simplicial depth is slightly better than that based on half-space depth.
- DD-classification can be based on other depth notions.
- For classification maximum depth was used here. But other classifications rules applied on the DD-plot can be used.

- DD-classification can be better than LDA, QDA, and RF.
- DD-classification based on simplicial depth is slightly better than that based on half-space depth.
- DD-classification can be based on other depth notions.
- For classification maximum depth was used here. But other classifications rules applied on the DD-plot can be used.
- Classification can be used for any dimension of data and also for functional data.

- DD-classification can be better than LDA, QDA, and RF.
- DD-classification based on simplicial depth is slightly better than that based on half-space depth.
- DD-classification can be based on other depth notions.
- For classification maximum depth was used here. But other classifications rules applied on the DD-plot can be used.
- Classification can be used for any dimension of data and also for functional data.
- Classification can be done also for more than two groups. But then a DD-Plot is not possible.

Liu, R.Y., Parelius, J.M., und Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by Liu and Singh). *Annals of Statistics* **27**, 783-858.

Zhang, Z., Cui, X., Jeske, D.R., und Borneman, J. (2013). Biclustering scatter plots using data depth measures. *Statistical Analysis and Data Mining* **6**, 102-115.

`http://cran.r-project.org/web/packages/depth/depth.pdf` (5.10.2015)

`https://cran.r-project.org/web/packages/fda.usc/fda.us` (5.10.2015)

`https://cran.r-project.org/web/packages/ddalpha/` (30.10.2015)