# Fast and Robust Classifiers Adjusted for Skewness

Mia Hubert and Stephan Van der Veeken

Katholieke Universiteit Leuven, Department of Mathematics

Mia.Hubert@wis.kuleuven.be

COMPSTAT 2010

# Outline

# Outline

- Review of some classifiers

# Outline

- Review of some classifiers
  - normally distributed data

# Outline

- Review of some classifiers
  - normally distributed data
  - depth based approaches

# Outline

- Review of some classifiers
  - normally distributed data
  - depth based approaches
- New approaches based on adjusted outyingness

# Outline



- Review of some classifiers
  - normally distributed data
  - depth based approaches
- New approaches based on adjusted outyingness
- Simulation results

# Outline



- Review of some classifiers
  - normally distributed data
  - depth based approaches
- New approaches based on adjusted outyingness
- Simulation results
- A real data set

# Outline



- Review of some classifiers
  - normally distributed data
  - depth based approaches
- New approaches based on adjusted outyingness
- Simulation results
- A real data set
- Conclusions and outlook

# Some classifiers



Setting:

- Observations sampled from $k$ different classes $X^j,\ j = 1, \ldots, k$.
- data belonging to group $X^j$ are denoted by $\boldsymbol{x}_i^j$ $(i = 1, \ldots, n_j)$
- the dimension of the data space is $p$ and $p \ll n_j$.
- outliers possible!

Classification:

construct a rule to classify a new observation into one of the $k$ populations.

# Some classifiers



Normally distributed data:

- Classical Linear discriminant analysis (when covariance matrices in each group are equal)
- Classical Quadratic discriminant analysis (CQDA)

based on classical mean and covariance matrices.

# Some classifiers



Normally distributed data:

- Classical Linear discriminant analysis (when covariance matrices in each group are equal)
- Classical Quadratic discriminant analysis (CQDA)

based on classical mean and covariance matrices.

Robust versions (RLDA, RQDA) are obtained by using robust covariance matrices, such as the MCD-estimator or S-estimators.

(He and Fung 2000, Croux and Dehon 2001, Hubert and Van Driessen 2004).

# Depth based classifiers



Proposed by Ghosh and Chaudhuri (2005).

- Consider a depth function (Tukey depth, simplicial depth, ...).
- For a new observation: compute its depth with respect to each group.
- Assign the new observation to the group for which it attains the **maximal depth**.

# Depth based classifiers



Advantages:

- does not rely on normality
- optimality results at normal data
- robust towards outliers (degree of robustness depends on depth function)
- can handle multigroup classification, not only two-group

# Depth based classifiers



Advantages:

- does not rely on normality
- optimality results at normal data
- robust towards outliers (degree of robustness depends on depth function)
- can handle multigroup classification, not only two-group

Disadvantages:

- computation time
- ties: observations outside the convex hull of all groups have zero depth w.r.t. each group
- adaptations necessary for unequal sample sizes. Ghosh and Chaudhuri propose methods that rely on kernel density estimates.

# New depth based classifiers



New proposals based on **adjusted outlyingness**.

First consider *univariate data*.

**Standard boxplot** has whiskers as the smallest and the largest data point that do not exceed:

$$[Q_1 - 1.5 \text{ IQR}, Q_3 + 1.5 \text{ IQR}]$$

**Adjusted boxplot** has whiskers that end at the smallest and the largest data point that do not exceed

$$[Q_1 - 1.5\, e^{-4 \text{ MC}} \text{ IQR}, Q_3 + 1.5\, e^{3 \text{ MC}} \text{ IQR}]$$

with

$$\text{MC}(X) = \underset{x_i < m < x_j}{\text{med}} h(x_i, x_j)$$

with $m$ the median of $X$ and

$$h(x_i, x_j) = \frac{(x_j - m) - (m - x_i)}{x_j - x_i}$$

(Hubert and Vandervieren, CSDA, 2008)

# Medcouple - A robust measure of skewness



- **Robustness**:
  - bounded influence function

    $\rightarrow$ adding a small probability mass at a certain point has a bounded influence on the estimate.

  - high breakdown point

    $$\epsilon^*(\mathrm{MC}) = 25\%$$

    $\rightarrow$ 25% of the data needs to be replaced to make the estimator break down

# Medcouple - A robust measure of skewness



- **Robustness**:
  - bounded influence function

    $\rightarrow$ adding a small probability mass at a certain point has a bounded influence on the estimate.

  - high breakdown point

    $$\epsilon^*(\mathrm{MC}) = 25\%$$

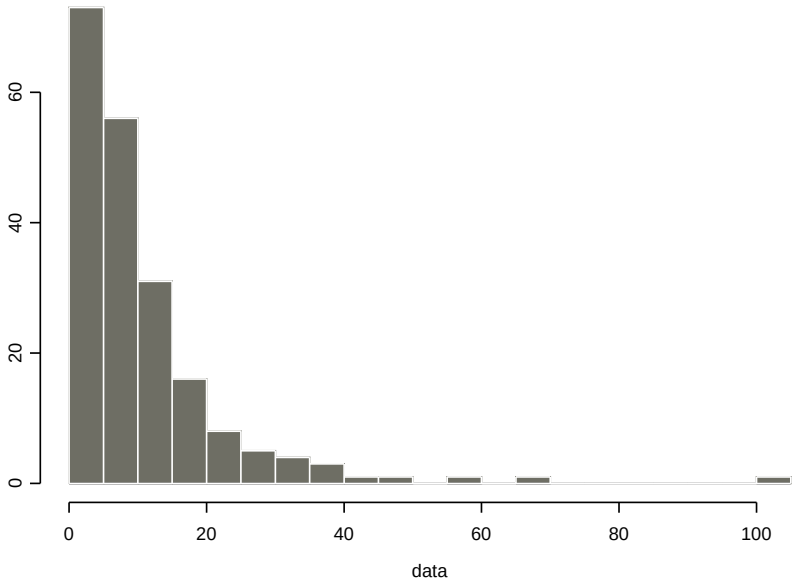
    $\rightarrow$ 25% of the data needs to be replaced to make the estimator break down

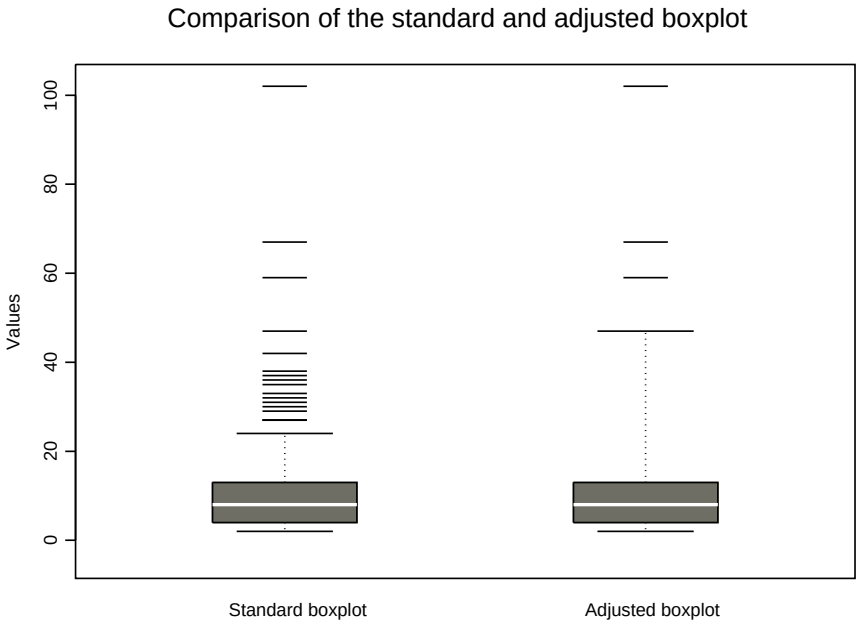
- **Computation**:
  - fast algorithm available O($n \log n$)

# Adjusted boxplot



## **Example:** Length of stay in hospital



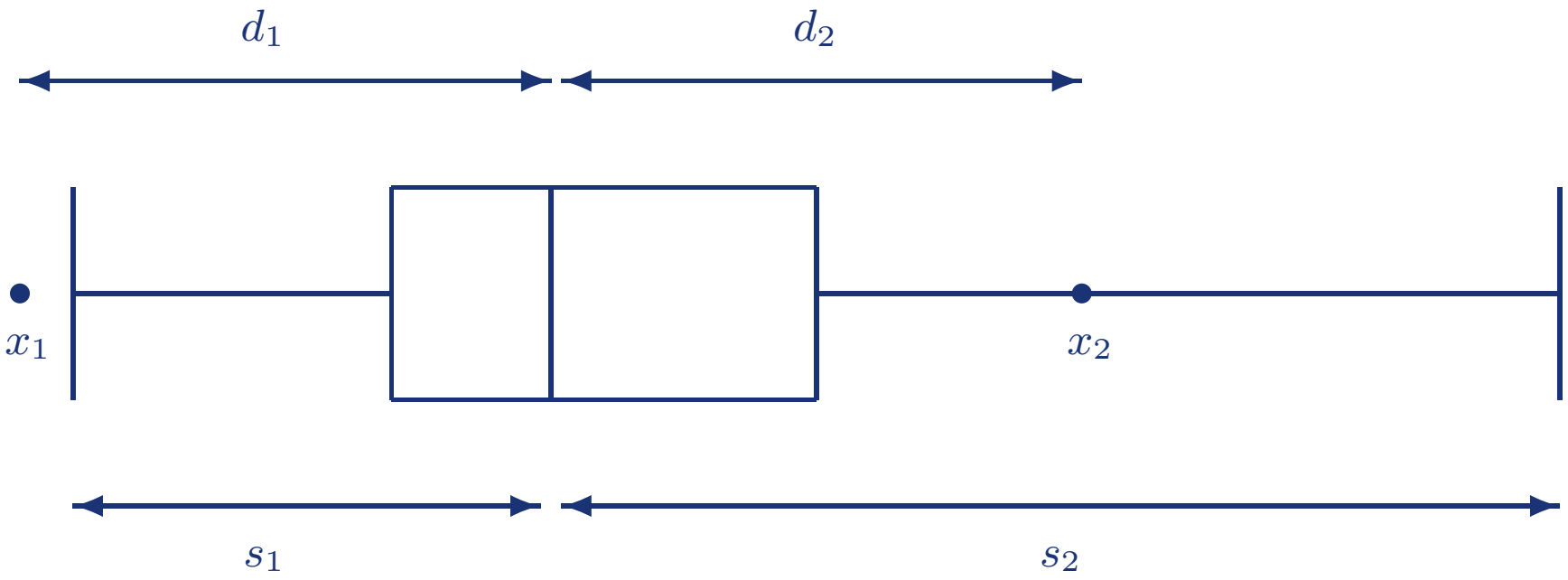
# Adjusted outlyingness - univariate data



For univariate data, the *adjusted outlyingness* is defined as:

$$\mathrm{AO}_i^{(1)} = \frac{|x_i - m|}{(w_2 - m)I[x_i > m] + (m - w_1)I[x_i < m]}$$

with $w_1$ and $w_2$ the whiskers of the adjusted boxplot.



# Adjusted outlyingness - univariate data



- $\text{AO}_i^{(1)}(x_1) = d_1/s_1$ and $\text{AO}_i^{(1)}(x_2) = d_2/s_2$.
- Although $x_1$ and $x_2$ are located at the same distance from the median, $x_1$ will have a higher value of adjusted outlyingness, because of the fact that the denominator $s_1$ is smaller.
- Skewness is thus used to estimate the scale differently on both sides of the median.
- Data-driven (outlying with respect to bulk of the data)

Brys, Hubert and Rousseeuw (2005), Hubert and Van der Veeken (2008)

# Adjusted outlyingness for multivariate data



Projection pursuit idea:

$$\mathrm{AO}_i = \mathrm{AO}(\boldsymbol{x}_i, X) = \sup_{\mathbf{a} \in \mathbb{R}^p} \mathrm{AO}^{(1)}(\mathbf{a}^t \boldsymbol{x}_i, X\mathbf{a}).$$

In practice:

consider $250p$ directions, generated as the direction perpendicular to the subspace spanned by $p$ observations, randomly drawn from the data set.

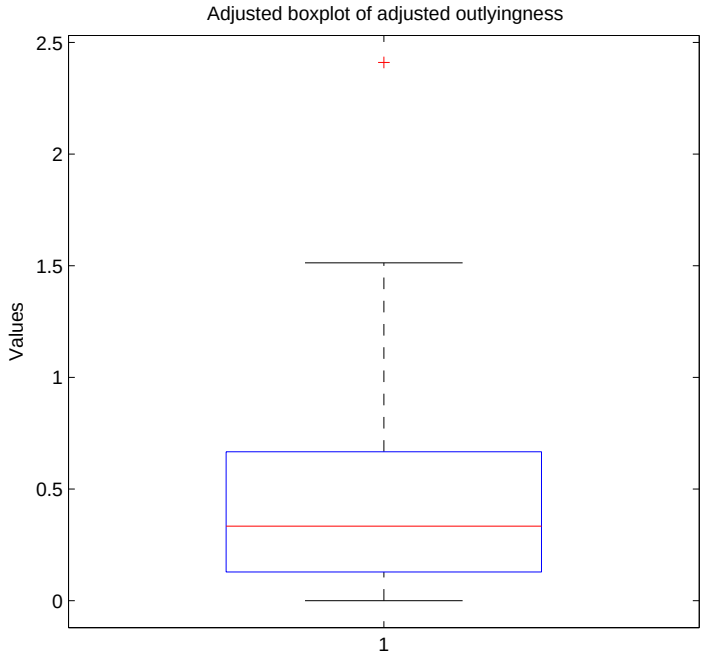
# Adjusted outlyingness



Outlier detection (for univariate as well as multivariate data):

- Construct adjusted boxplot of the $AO_i$
- Outliers: observations whose $AO_i$ exceeds the upper whisker

**Example**: Length of stay, $n = 201$



# Depth classifier - minimal AO



Classifier 1:

> Assign the new observation to the group for $\mathrm{AO}(\boldsymbol{y}, X^j)$ is minimal.

Hubert and Van der Veeken (2010)

Related to projection depth :

$$\mathrm{PD}(\boldsymbol{x}_i, X) = 1/(1 + \mathrm{O}(\boldsymbol{x}_i, X))$$

with $\mathrm{O}(\boldsymbol{x}_i, X)$ the Stahel-Donoho outlyingness (which does not use a skewness estimate)

(Zuo and Serfling 2000, Dutta and Ghosh 2009, Cui et al. 2008)

# Depth classifier - minimal AO



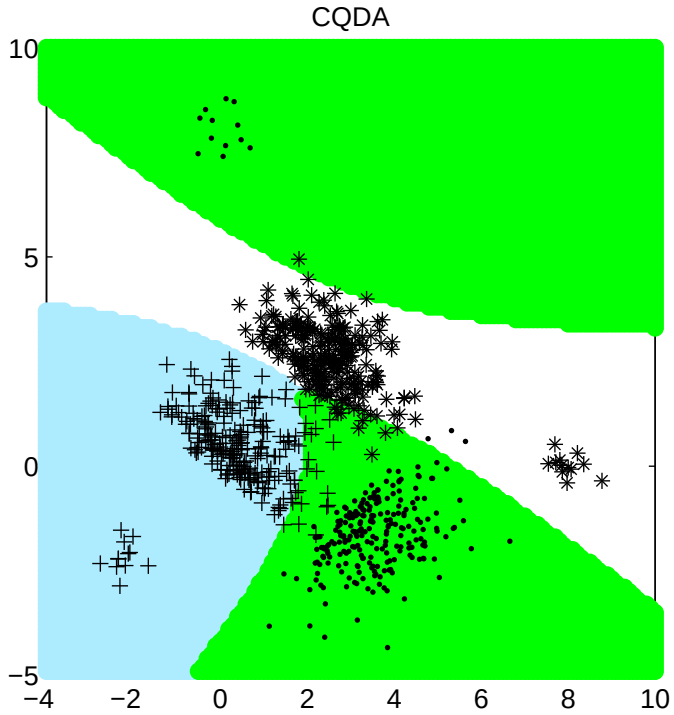
More precisely:

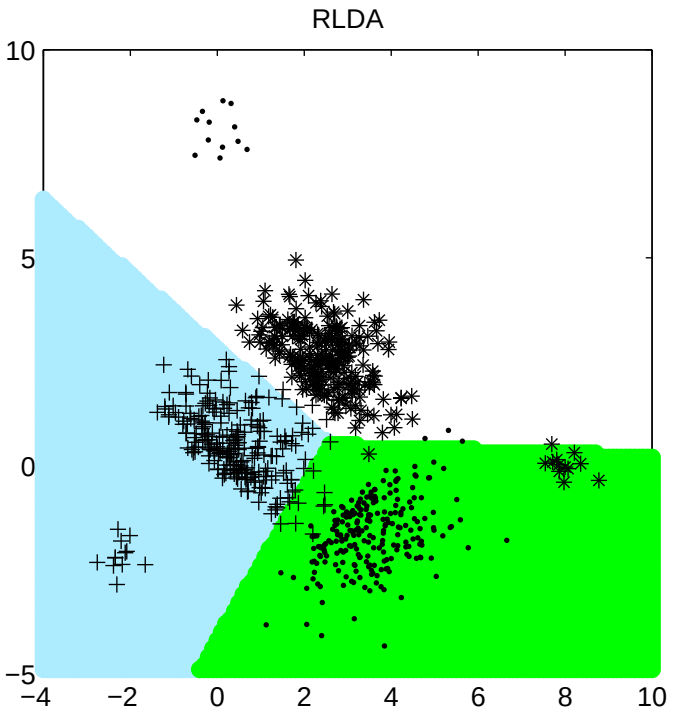
- First compute the $\text{AO}_i^j(\boldsymbol{x}_i^j, X^j)$ (outlyingness of all observations from group $j$ w.r.t. $X^j$)
- Remove outliers from $X^j$ based on these $\text{AO}_i^j$. This yields $\tilde{X}^j$ with sample size $\tilde{n}_j$.
- Recompute the $\text{AO}_i^j(\boldsymbol{x}_i^j, \tilde{X}^j)$ for all $\boldsymbol{x}_i^j$ in $\tilde{X}^j$. This gives $\{\tilde{\text{AO}}^j\}$. Retain median, mad, MC computed in each direction.
- For a new observation $\boldsymbol{y}$, compute $\text{AO}(\boldsymbol{y}, \tilde{X}^j)$ based on the medians, mads, MCs from previous step.

# Depth classifier - minimal AO

Illustration: three groups generated from skew-normal distributions.

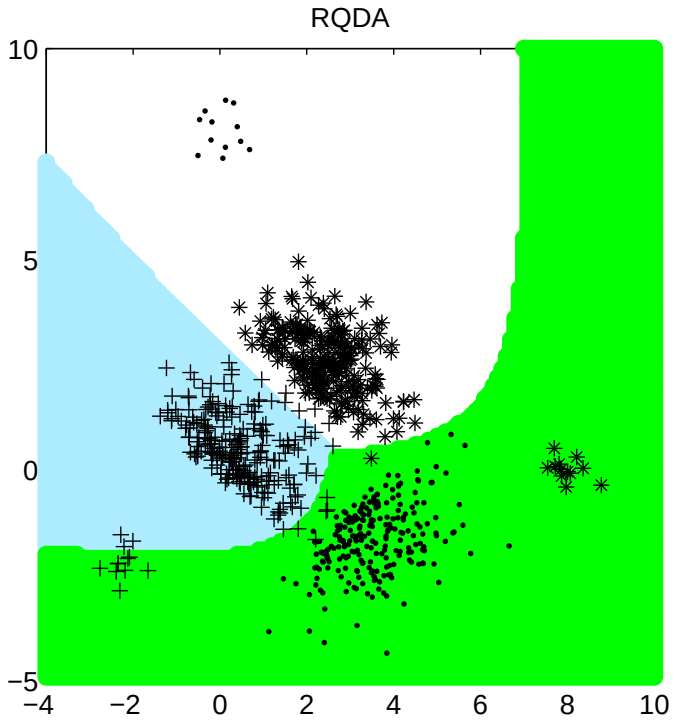


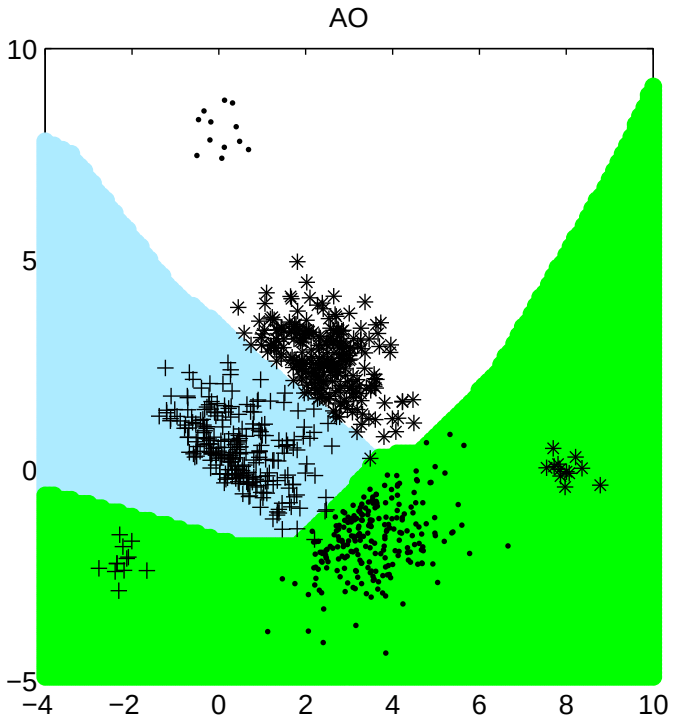


# Depth classifier - minimal AO





# Depth classifier - minimal AO



Some simulation results:

- $n_j$ training data generated from *three* **skew-normal** distributions
- $p = 2$ then $n_j = 250$
  $p = 3$ and $p = 5$ then $n_j = 500$
- also outliers introduced
- test data $n_j/5$ from same distributions
- misclassification errors of the test set (average and standard errors over 100 simulations)
- comparison with CQDA, RLDA and RQDA based on the MCD-estimator

# Depth classifier - minimal AO



| | $\varepsilon$ | CQDA | RLDA | RQDA | AO |
|---|---|---|---|---|---|
| 2D | 0% | 0.0234 | 0.0254 | 0.0193 | 0.0117 |
| | | (0.001) | (0.0011) | (0.0012) | (0.0012) |
| | 5% | 0.0341 | 0.0228 | 0.0170 | 0.0127 |
| | | (0.0015) | (0.0013) | (0.0011) | (0.0011) |
| 3D | 0% | 0.0228 | 0.0240 | 0.0191 | 0.0120 |
| | | (0.0006) | (0.0008) | (0.0008) | (0.0008) |
| | 5% | 0.0304 | 0.0209 | 0.0181 | 0.0127 |
| | | (0.001) | (0.0006) | (0.0006) | (0.0007) |
| 5D | 0% | 0.0125 | 0.0135 | 0.0141 | 0.0106 |
| | | (0.0006) | (0.0008) | (0.0007) | (0.0007) |
| | 5% | 0.0179 | 0.0140 | 0.0144 | 0.0114 |
| | | (0.0008) | (0.0008) | (0.0008) | (0.0007) |

# Depth classifier - minimal AO



Simulation results for elliptical data:

- $n_j$ training data generated from *two* **normal** distributions
- $p = 2$ then $n_j = 250$
  $p = 3$ and $p = 5$ then $n_j = 500$
- also outliers introduced
- test data $n_j/5$ from same distributions
- misclassification errors of the test set (average and standard errors over 100 simulations)
- comparison with CQDA, RLDA, RQDA and LS-SVM with RBF kernel

# Depth classifier - minimal AO



| | $\varepsilon$ | CQDA | RLDA | RQDA | AO | LS-SVM |
|---|---|---|---|---|---|---|
| 2D | 0% | 0.0763 | 0.0762 | 0.0777 | 0.0821 | 0.0801 |
| | | (0.0028) | (0.0027) | (0.0026) | (0.0029) | (0.0024) |
| | 10% | 0.1545 | 0.0808 | 0.0795 | 0.0839 | 0.0825 |
| | | (0.0052) | (0.0026) | (0.0026) | (0.0026) | (0.0025) |
| 3D | 0% | 0.0421 | 0.0426 | 0.0430 | 0.0448 | 0.0435 |
| | | (0.0015) | (0.0014) | (0.0014) | (0.0015) | (0.0015) |
| | 10% | 0.1327 | 0.0432 | 0.0429 | 0.0452 | 0.0430 |
| | | (0.0036) | (0.0014) | (0.0014) | (0.0014) | (0.0014) |
| 5D | 0% | 0.1310 | 0.1308 | 0.1325 | 0.1465 | 0.1339 |
| | | (0.0025) | (0.0025) | (0.0024) | (0.0026) | (0.0024) |
| | 10% | 0.2122 | 0.1340 | 0.1363 | 0.1572 | 0.1390 |
| | | (0.0038) | (0.0025) | (0.0025) | (0.0025) | (0.0025) |

# Adjustments for unequal group sizes



Inspired by Billor et al. (2008): assign the new observation to the group for which its depth has *highest rank*.

Classifier 2:

Let $r_y^j$ be the distribution function of $\mathrm{AO}(\boldsymbol{y}, \tilde{X}^j)$ with respect to the $\{\tilde{\mathrm{AO}}^j\}$:

$$r_y^j = \frac{1}{\tilde{n}_j} \sum_{i=1}^{\tilde{n}_j} I(\tilde{\mathrm{AO}}_i^j \leqslant \mathrm{AO}(\boldsymbol{y}, \tilde{X}^j)).$$

Assign observation $\boldsymbol{y}$ to the group $j$ for which $r_y^j$ is minimal.

(If ties, then use classifier 1.)

The e.d.f. is a way to measure the position of $\mathrm{AO}(\boldsymbol{y}, \tilde{X}^j)$ within the $\{\tilde{\mathrm{AO}}^j\}$.

# Adjustments for unequal group sizes



Classifier 3:

To measure the position of $\mathrm{AO}(\boldsymbol{y}, \tilde{X}^j)$, we use a distance which is related to the definition of the univariate AO. Let in general

$$\mathrm{SAO}^{(1)}(x, X) = \mathrm{AO}^{(1)}(x, X)\ \mathrm{sign}(x - \mathrm{med}(X))$$

be the *signed* adjusted outlyingness of an observation $x$ with respect to a univariate data set $X$.

Let

$$s_y^j = \mathrm{SAO}^{(1)}(\mathrm{AO}(\boldsymbol{y}, \tilde{X}^j), \{\tilde{\mathrm{AO}}^j\}).$$

Assign observation $\boldsymbol{y}$ to the group $j$ for which $s_y^j$ is minimal.

# Adjustments for unequal group sizes



Simulation results for equal sample sizes: $n_1 = n_2 = 500$:

| | $\varepsilon$ | Classifier 1 | Classifier 2 | Classifier 3 |
|---|---|---|---|---|
| 2D | 0% | 0.0737 | 0.0751 | 0.0758 |
| | | (0.0018) | (0.0019) | (0.0019) |
| | 5% | 0.0744 | 0.0751 | 0.0756 |
| | | (0.0021) | (0.0021) | (0.0021) |
| 3D | 0% | 0.0440 | 0.0449 | 0.0451 |
| | | (0.0015) | (0.0016) | (0.0016) |
| | 5% | 0.0425 | 0.0437 | 0.0425 |
| | | (0.0015) | (0.0015) | (0.0015) |
| 5D | 0% | 0.0737 | 0.0749 | 0.0758 |
| | | (0.0015) | (0.0017) | (0.0018) |
| | 5% | 0.0736 | 0.0735 | 0.0767 |
| | | (0.0016) | (0.0016) | (0.0019) |

# Adjustments for unequal group sizes



Simulation results for unequal sample sizes: $n_1 = 100$ and $n_2 = 500$

| | $\varepsilon$ | Classifier 1 | Classifier 2 | Classifier 3 |
|---|---|---|---|---|
| 2D | 0% | 0.1047 | 0.0882 | 0.0876 |
| | | (0.0033) | (0.0026) | (0.0026) |
| | 5% | 0.0991 | 0.0797 | 0.0818 |
| | | (0.0032) | (0.0024) | (0.0023) |
| 3D | 0% | 0.0986 | 0.0527 | 0.0534 |
| | | (0.0032) | (0.0015) | (0.0015) |
| | 5% | 0.0965 | 0.0533 | 0.0499 |
| | | (0.0032) | (0.0018) | (0.0017) |
| 5D | 0% | 0.2298 | 0.0930 | 0.0909 |
| | | (0.0042) | (0.0026) | (0.0028) |
| | 5% | 0.2284 | 0.0956 | 0.0916 |
| | | (0.0041) | (0.0023) | (0.0028) |

# Example



Data from the Belgian Household Survey of 2005.

$X_1$ : Income
$X_2$ : Expenditure on durable consumer goods.

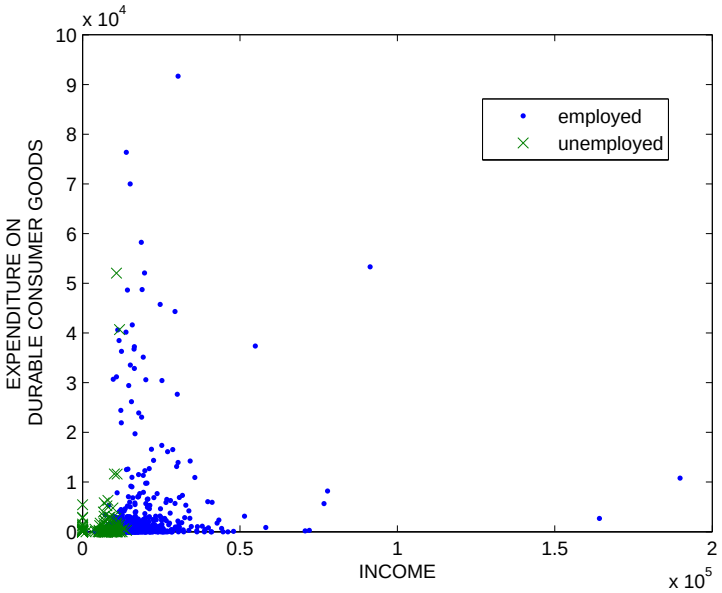
To avoid correcting factors for family size, only single persons are considered.

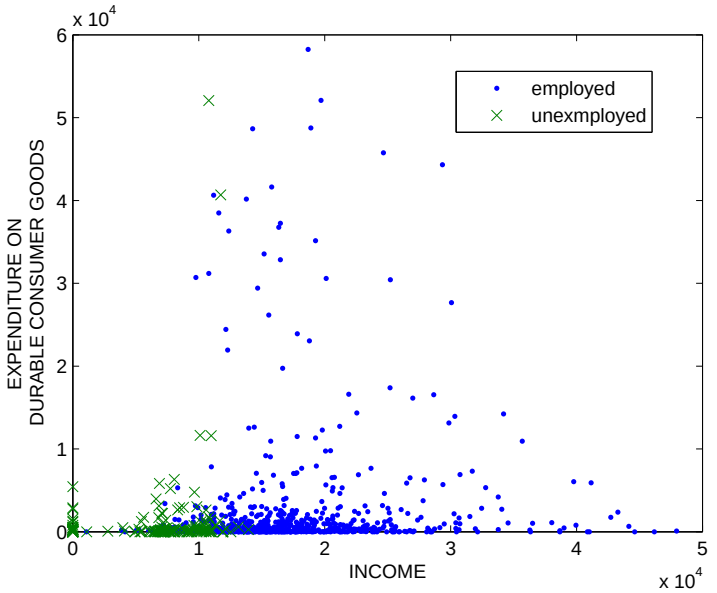
This group of single persons consists of 174 *unemployed* and 706 (at least partially) *employed* persons.

Goal: classification of a person as employed or unemployed based on income and expenditure on durable consumer goods.

# Example






Both groups are randomly split into a training and a test set which contains 10 data points.

Average misclassification errors (over 100 replications):

Classifier 1: 0.2580 (s.e. 0.0099)
Classifier 2: 0.1655 (s.e. 0.0082)
Classifier 3: 0.1855 (s.e. 0.0086).

# Conclusion and outlook



- Classifiers that adjust for skewness and sample sizes yield lower misclassification errors
- Classifiers can be computed fast in any dimension (depends on number of directions considered)
- Could also be used in the DD-plot (depth-versus-depth plot) proposed by Li et al. (2010).
- Programs soon available in LIBRA, Matlab LIBrary for Robust Analysis at

  `wis.kuleuven.be/stat/robust`

- Extensions available for high-dimensional data: combining robust PCA for skewed data and RSIMCA.

# Some references




- Hubert, M. and Van der Veeken, S. (2008). Outlier detection for skewed data. *Journal of Chemometrics* 22, 235–246.
- Hubert, M., and Van der Veeken, S. (2010). Robust classification for skewed data. *Advances in Data Analysis and Classification*, in press.
- Hubert, M., and Van der Veeken, S. (2010). Fast and robust classifiers adjusted for skewness. *Proceedings of Compstat 2010*.
- Billor, N., Abebe, A., Turkmen, A. and Nudurupati, S.V. (2008). Classification based on depth transvariations. *Journal of Classification* 25, 249-260.
- Dutta, S, Ghosh, A.K. (2009). On robust classification using projection depth. *Indian Statistical Institute,* Technical report R11/2009.
- Ghosh, A.K., and Chaudhuri, P. (2005). On maximum depth and related classifiers. *Scandinavian Journal of Statistics* 32, 327–350.
- Li, J., Cuesta-Albertos, J.A., Liu, R.Y. (2010). DD-classifier: nonparamatric classification procedure based on DD-plot. Submitted.