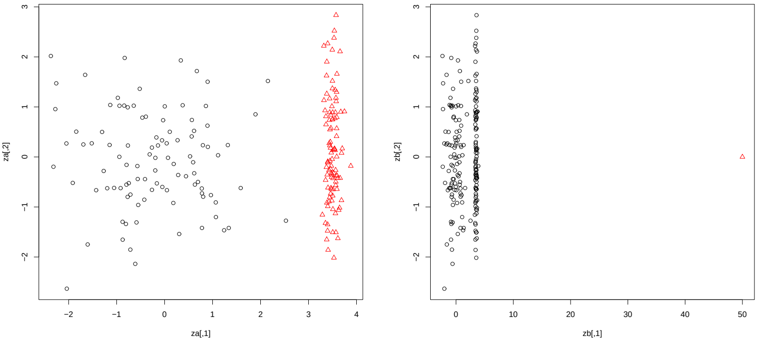
Some thoughts about robust clustering

Christian Hennig

October 5, 2009

1. Introduction

Nonrobustness of mean/covariance matrix-based methods well known.



Done by Gaussian mixture ML clustering:

$$f(x) = \sum_{i=1}^{s} \pi_i \varphi_{\mathbf{a}_i, \Sigma_i}(x).$$

Done by Gaussian mixture ML clustering:

$$f(x) = \sum_{i=1}^{s} \pi_i \varphi_{\mathbf{a}_i,\Sigma_i}(x).$$

Number of components can be estimated by BIC ($s = 2$ here).
Various covariance matrix models (free, all equal, spherical. . .).
Given the estimated parameters, points can be classified by max. $\hat{P}(\gamma = j|x) = (\hat{\pi}_j \varphi_{\hat{\mathbf{a}}_j,\hat{\Sigma}_j}(x))/(\sum_{i=1}^{k} \hat{\pi}_i \varphi_{\hat{\mathbf{a}}_i,\hat{\Sigma}_i}(x))$
Software R-package mclust (Fraley and Raftery).

Similar robustness problems for k-means, other methods.

2. Approaches to robust clustering

2.1 Mixture model

- Mixtures of t-distributions (McLachlan & Peel 2000)

2. Approaches to robust clustering

2.1 Mixture model

- Mixtures of t-distributions (McLachlan & Peel 2000)
- The "noise component" (Banfield & Raftery, 1993)

$$f(x) = \pi_0 \frac{1}{V} + \sum_{j=1}^{s} \pi_j \varphi_{\mathbf{a}_j, \Sigma_j}(x),$$

2. Approaches to robust clustering

2.1 Mixture model

- Mixtures of t-distributions (McLachlan & Peel 2000)
- The “noise component” (Banfield & Raftery, 1993)

$$f(x) = \pi_0 \frac{1}{V} + \sum_{j=1}^{s} \pi_j \varphi_{\mathbf{a}_j,\Sigma_j}(x),$$

- Improper noise (Hennig 2004, Coretto 2008)

$$f(x) = \pi_0 c + \sum_{j=1}^{s} \pi_j \varphi_{\mathbf{a}_j,\Sigma_j}(x),$$

2. Approaches to robust clustering

2.1 Mixture model

- Mixtures of t-distributions (McLachlan & Peel 2000)
- The “noise component” (Banfield & Raftery, 1993)

$$f(x) = \pi_0 \frac{1}{V} + \sum_{j=1}^{s} \pi_j \varphi_{\mathbf{a}_j,\Sigma_j}(x),$$

- Improper noise (Hennig 2004, Coretto 2008)

$$f(x) = \pi_0 c + \sum_{j=1}^{s} \pi_j \varphi_{\mathbf{a}_j,\Sigma_j}(x),$$

- Trimmed likelihood (Neykov, Filzmoser, Dimova, Neytchev 2007)

2.2 Fixed Partition Models

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} \varphi_{\mathbf{a}_{\gamma(i)}, \Sigma_{\gamma(i)}}(x_i),$$

$\gamma : \{1, \ldots, n\} \rightarrow \{1, \ldots, s\}$.
k-means is a fixed partition ML-method
(equal spherical covariance matrices).

2.2 Fixed Partition Models

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} \varphi_{\mathbf{a}_{\gamma(i)}, \Sigma_{\gamma(i)}}(x_i),$$

$\gamma : \{1, \ldots, n\} \to \{1, \ldots, s\}$.
k-means is a fixed partition ML-method
(equal spherical covariance matrices).
Crisp assignment of points,
inconsistent for parameter estimation,
but often good for classification.

Robustification:

- Replacing mean/covariance matrix by robust estimators (medians, MCD)
  (Kaufman & Rousseeuw 1990, Rocke & Hardin 2000)

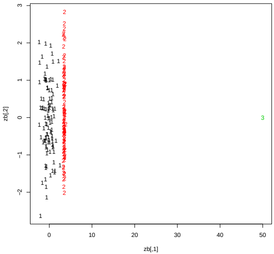
Robustification:

- Replacing mean/covariance matrix by robust estimators (medians, MCD)
(Kaufman & Rousseeuw 1990, Rocke & Hardin 2000)
- **Trimming:** optimise ML-criterion for best $n - r$ points.
α-trimmed k-means
(Cuesta-Albertos, Gordaliza, Matran 1997)
- Equal covariance matrices (Gallegos & Ritter 2005)
- Flexible covariance matrices
(Garcia-Escudero, Gordaliza, Matran, Mayo-Iscar 2008)

2.3 Finding outlier-free data subsets

- Fixed point clusters (Hennig & Christlieb 2002)
- Forward search (Atkinson, Cerioli & Riani 2003)

How the methods do (with $s = 2$)



- Outlier only "noise point" (proper choice of c)
- t-mixtures integrate it with cluster 1.
- Trimming: will be trimmed.
  Need flexible covarince matrices, proper trimming rate.

3. Breakdown and robustness measurement
3.1 Breakdown theory for mixtures (Hennig 2004)

Finite sample addition **breakdown point**
(Donoho and Huber 1983):
smallest possible contamination ($\frac{g}{n+g}$)
to be added to the dataset
so that estimator becomes “arbitrarily bad”.

Definition 1. Be $\hat{\theta}_{n,s}: \mathbf{x}_n \rightarrow \theta \in \Theta$ estimator, where

$$\Theta = \{(\pi_1, \ldots, \pi_s, a_1, \ldots, a_s, \sigma_1, \ldots, \sigma_s) : \sum \pi_j = 1,\ \sigma_j \geq c > 0 \forall j\}.$$

For $g > 0$ let

$$Y_g(\mathbf{x}_n) = \{\mathbf{x}_{n+g} \in \mathbb{R}^{n+g} : \text{ first n observations equal } \mathbf{x}_n\}$$

Breakdown point $\hat{\theta}_{\bullet,s}$: $B(\hat{\theta}_{\bullet,s}, \mathbf{x}_n) = \min \frac{g}{n+g}$ so that at least one of the following for at least one $i \in \{1, \ldots, s\}$:

- $\inf_{\mathbf{x}_{n+g} \in Y_g(\mathbf{x}_n)} \hat{\pi}_i(\mathbf{x}_{n+g}) = 0$ for i with $\hat{\pi}_i(\mathbf{x}_n) > 0$,
- $\sup_{\mathbf{x}_{n+g} \in Y_g(\mathbf{x}_n)} \hat{\sigma}_i(\mathbf{x}_{n+g}) = \infty$,
- $\sup_{\mathbf{x}_{n+g} \in Y_g(\mathbf{x}_n)} |\hat{\mathbf{a}}_i(\mathbf{x}_{n+g})| = \infty$.

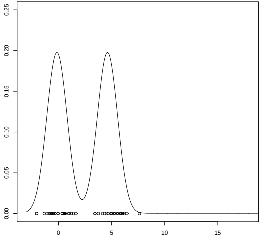
Theorem 2. *Let $s > 1$,*

$$\hat{\theta}_{n,s}(\mathbf{x}_n) = \arg\max_{\theta \in \Theta} L_{n,s}(\theta, \mathbf{x}_n),$$

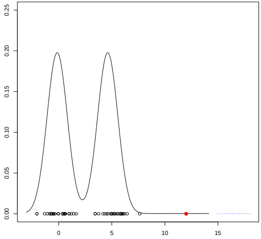
$$L_{n,s}(\theta, \mathbf{x}_n) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{s} \pi_j \varphi_{\mathbf{a}_j,\sigma_j}(x_i)\right),$$

the ML-estimator. Then $B(\hat{\theta}_{\bullet,s}, \mathbf{x}_n) = \frac{1}{n+1}$.

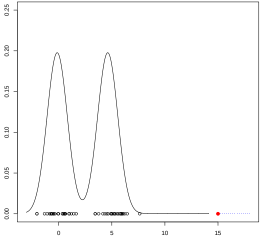
$$L_{n,s}(\eta) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{s} \pi_j f_{a_j,\sigma_j}(x_i)\right) + \log\left(\sum_{j=1}^{s} \pi_j f_{a_j,\sigma_j}(x_{n+1})\right).$$



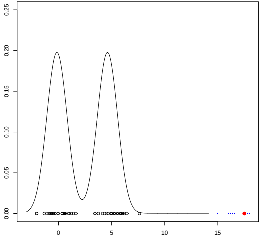
$$L_{n,s}(\eta) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{s} \pi_j f_{a_j,\sigma_j}(x_i)\right) + \log\left(\sum_{j=1}^{s} \pi_j f_{a_j,\sigma_j}(x_{n+1})\right).$$



$$L_{n,s}(\eta) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{s} \pi_j f_{a_j,\sigma_j}(x_i)\right) + \log\left(\sum_{j=1}^{s} \pi_j f_{a_j,\sigma_j}(x_{n+1})\right).$$

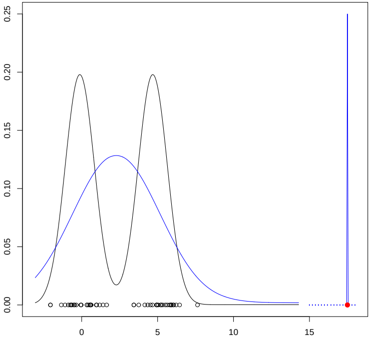


$$L_{n,s}(\eta) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{s} \pi_j f_{a_j,\sigma_j}(x_i)\right) + \log\left(\sum_{j=1}^{s} \pi_j f_{a_j,\sigma_j}(x_{n+1})\right).$$



No breakdown of original components:

$$x_{n+1} \to \infty \Rightarrow L_{n,s} \to -\infty$$



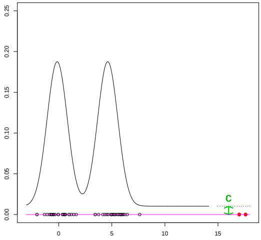
$$a_s = x_{n+1} \Rightarrow L_{n,s} \geq \text{const, thus } B = \frac{1}{n+1}$$

(as well for t-mixtures and $\frac{1}{V}$-noise component)

Improper noise:

$$\hat{\eta}_{n,s} = \arg\max_{\eta} \sum_{i=1}^{n} \log\left(\sum_{j=1}^{s} \pi_j f_{a_j,\sigma_j}(x_i) + \pi_0 c\right)$$

with fixed c

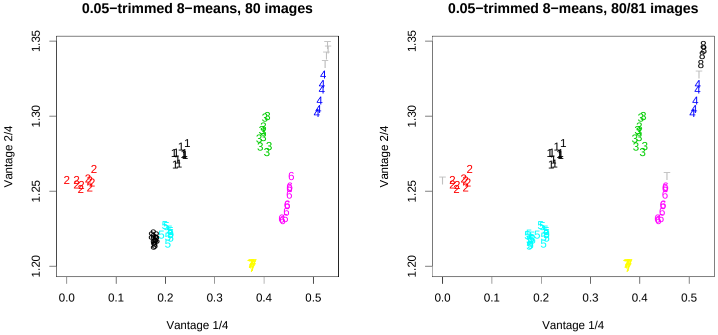


Theorem 3.
$\forall r < s: \; (\hat{L}_{n,s} - \hat{L}_{n,r}) >$
$n \log \frac{cg}{n} + g \log((\pi_0 + \frac{g}{n})c) + (n+g) \log \frac{n}{n+g} - g \log f_{max}$
$\Rightarrow B_n > \frac{g}{n+g}$ (data dependent)

Theorem 3.
$\forall r < s: \ (\hat{L}_{n,s} - \hat{L}_{n,r}) >$
$n \log \frac{cg}{n} + g \log((\pi_0 + \frac{g}{n})c) + (n+g) \log \frac{n}{n+g} - g \log f_{max}$
$\Rightarrow B_n > \frac{g}{n+g}$ (data dependent)

Trimming: Gallegos & Ritter (2005)
have data dependent breakdown point
under separation condition, too.
Garcia-Escudero & Gordaliza (1999)
know that it's data dependent.

3.2 The dissolution point (Hennig 2008)



Clusters may break down without parameters breaking down.
Some clustering methods don't estimate parameters.

Aim: “something like” a breakdown point
for general clustering methods (GCM):

$$E = (E_n)_{n \in \mathbb{N}},$$
$$E_n : \mathbf{x}_n \mapsto \{C_1, \ldots, C_k\}, \; C_j \subseteq \mathbf{x}_n.$$

(Assume $C_i \cap C_j = \emptyset$.)

Wanted:
results in terms of set memberships,
not parameters.

Induced clustering (g points added):

$$E_n^*(\mathbf{x}_{n+g}) = E_{n+g}(\mathbf{x}_{n+g}) \cap \mathbf{x}_n$$

Cluster similarity (Jaccard, 1901):

$$\gamma(C,D) = \frac{|C \cap D|}{|C \cup D|},\ \gamma^*(C,\mathcal{D}) = \max_{D \in \mathcal{D}} \gamma(C,D).$$

Principle: A cluster is **dissolved**
if the closest cluster in induced clustering (under addition)
is too far away. (How far is "too far"?)

Definition 4.

$$\Delta(E, \mathbf{x}_n, C) = \min_g \left\{ \frac{g}{|C|+g} : \exists \mathbf{x}_{n+g} = (x_1, \ldots, x_{n+g}) : \gamma^*(C, E_n^*(\mathbf{x}_{n+g})) \leq \frac{1}{2} \right\}$$

is called **dissolution point** of cluster C.

1/2 not worst possible value, but. . .

Definition 4.

$$\Delta(E, \mathbf{x}_n, C) = \min_g \left\{ \frac{g}{|C|+g} : \ \exists \mathbf{x}_{n+g} = (x_1, \ldots, x_{n+g}) : \gamma^*(C, E_n^*(\mathbf{x}_{n+g})) \leq \frac{1}{2} \right\}$$

is called **dissolution point** of cluster C.

1/2 not worst possible value, but...

- $\frac{1}{2}$ is minimal such that
  - for $n \geq 2$, $k \geq 2$, every cluster can dissolve,
  - whenever $|\mathcal{C}| = s$, $|\mathcal{D}^*| = s - r$, at least r members of $\mathcal{C}$ dissolve.

Example of a dissolution result

Theorem 5. Let E_k be α-trimmed k-means clustering.
For $g \leq n - \lceil n(1-\alpha) \rceil$,
if for any possible induced clustering $\mathcal{C}^*$ leading to dissolution of C:

$$\min_{y_1,\ldots,y_g \in D(E_{k,n}(\mathbf{x}_n))} \sum_{i=1}^{g} \min_j \|y_i - \bar{x}_j\|_2^2 < Q(\mathbf{x}_n, \mathcal{C}^*) - Q(\mathbf{x}_n, E_{k,n}(\mathbf{x}_n)),$$

then $\Delta(E_k, \mathbf{x}_n, C) > \frac{g}{|C|+g}$.

4. **Estimated** s (by BIC; Hennig 2004)

Definition 6. Let $\hat{\theta}_n : \mathbf{x}_n \to \theta \in \Theta_*$, Θ_* as Θ in Definition 1 but with parameter $s \in \mathbb{N}$.
The **breakdown point** of estimator $\hat{\theta}_\bullet$ is $B(\hat{\theta}_\bullet, \mathbf{x}_n) = \min \frac{g}{n+g}$ so that

$\inf_{\mathbf{x}_{n+g} \in Y_g(\mathbf{x}_n)} \hat{s}(\mathbf{x}_{n+g}) < \hat{s}(\mathbf{x}_n)$,
or at least one of the following for at least one $i \in \{1, \ldots, \hat{s}(\mathbf{x}_n)\}$:

- $\inf_{\mathbf{x}_{n+g} \in Y_g(\mathbf{x}_n)} \hat{\pi}_i(\mathbf{x}_{n+g}) = 0$ for i with $\hat{\pi}_i(\mathbf{x}_n) > 0$,
- $\sup_{\mathbf{x}_{n+g} \in Y_g(\mathbf{x}_n)} \hat{\sigma}_i(\mathbf{x}_{n+g}) = \infty$,
- $\sup_{\mathbf{x}_{n+g} \in Y_g(\mathbf{x}_n)} |\hat{\mathbf{a}}_i(\mathbf{x}_{n+g})| = \infty$.

Theorem 7. *Let $\hat{\theta}_{n,s}$ be the ML-estimator as before, and $\hat{\theta}_n(\mathbf{x}_n) = (\hat{s}(\mathbf{x}_n), \hat{\theta}_{n,\hat{s}}(\mathbf{x}_n))$ where $\hat{s}(\mathbf{x}_n)$ is the optimal number of components according to the BIC.*
Then $B(\hat{\theta}_\bullet, \mathbf{x}) \geq \frac{g}{n+g}$ for $\mathbf{x}_n$ so that

$$\min_{r<\hat{s}} L_{n,r}(\theta_{n,r}(\mathbf{x}_n), \mathbf{x}_n) - L_{n,s}(\theta_{n,\hat{s}}(\mathbf{x}_n), \mathbf{x}_n) > f(g),$$

f monotonically increasing positive finite function of g.

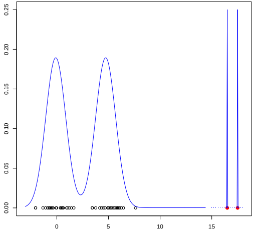
Theorem 7. *Let $\hat{\theta}_{n,s}$ be the ML-estimator as before, and $\hat{\theta}_n(\mathbf{x}_n) = (\hat{s}(\mathbf{x}_n), \hat{\theta}_{n,\hat{s}}(\mathbf{x}_n))$ where $\hat{s}(\mathbf{x}_n)$ is the optimal number of components according to the BIC.*
Then $B(\hat{\theta}_\bullet, \mathbf{x}) \geq \frac{g}{n+g}$ for $\mathbf{x}_n$ so that

$$\min_{r<\hat{s}} L_{n,r}(\theta_{n,r}(\mathbf{x}_n), \mathbf{x}_n) - L_{n,s}(\theta_{n,\hat{s}}(\mathbf{x}_n), \mathbf{x}_n) > f(g),$$

f monotonically increasing positive finite function of g.

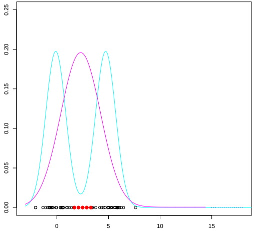
Analogous result for dissolution point.

$$\mathrm{BIC}(s) = \sum_{i=1}^{n+g} \log\left(\sum_{j=1}^{s} \pi_j f_{a_j,\sigma_j}(x_i)\right) - (3s-1)\log n.$$

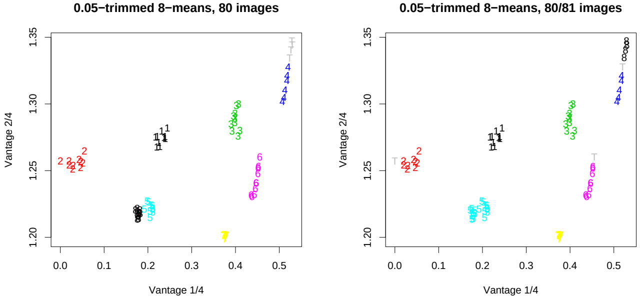


Breakdown by *outliers* almost impossible!

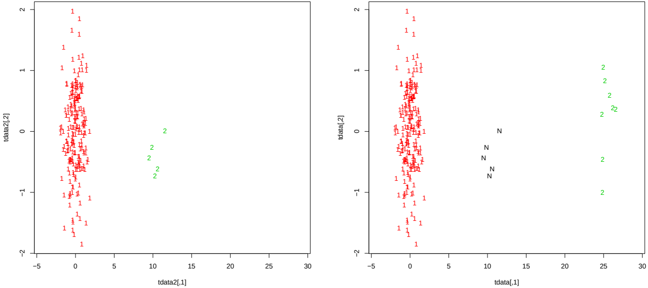
$$\mathrm{BIC}(s) = \sum_{i=1}^{n+g} \log \left(\sum_{j=1}^{s} \pi_j f_{a_j,\sigma_j}(x_i) \right) - (3s-1)\log n.$$



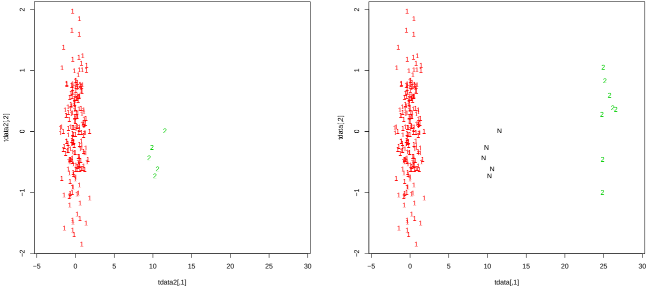
. . . but “in-between-liers” may cause trouble.
May still often be unstable.



Unclear $s \Rightarrow$
any fixed s method breaks down.



Add more points than size of smallest cluster:
any fixed s method breaks down (on any dataset).



Add more points than size of smallest cluster:
any fixed *s* method breaks down (on any dataset).
Estimated *s*: add clusters to avoid these problems.

So do we just have to estimate s then?
(...and not worry about noise, trimming etc.?)

So do we just have to estimate s then?

(. . . and not worry about noise, trimming etc.?)

- May have to fit lots of clusters if many isolated outliers are present.
  Computation & interpretation difficult.

So do we just have to estimate s then?
(...and not worry about noise, trimming etc.?)

- May have to fit lots of clusters if many isolated outliers are present.
  Computation & interpretation difficult.
- May rather want points classified as “noise” for reasons of interpretation.

So do we just have to estimate s then?
(...and not worry about noise, trimming etc.?)

- May have to fit lots of clusters if many isolated outliers are present.
  Computation & interpretation difficult.
- May rather want points classified as “noise” for reasons of interpretation.

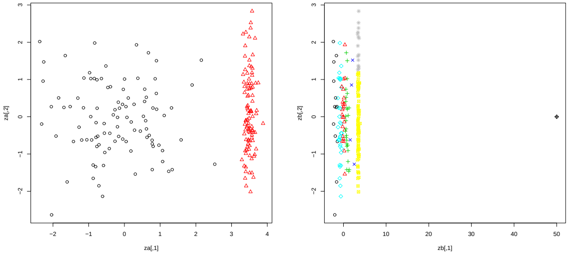
Open problem: compare (practical) robustness between clustering methods estimating s.

So do we just have to estimate s then?
(...and not worry about noise, trimming etc.?)

- May have to fit lots of clusters if many isolated outliers are present.
  Computation & interpretation difficult.
- May rather want points classified as "noise" for reasons of interpretation.

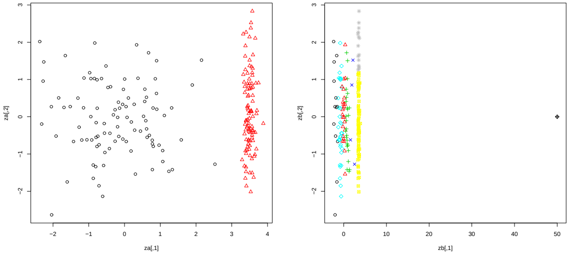
Open problem: compare (practical) robustness between clustering methods estimating s.
(Can do that data dependently by bootstrap methods; Hennig 2007)

"Death by flexibility"



mclustBIC solution.

"Death by flexibility"



mclustBIC solution.

No lower bound for cov-eigenvalue specified in mclustBIC.

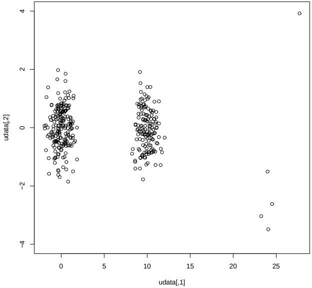
One-point cluster degenerates L with flexible cov-matrix.

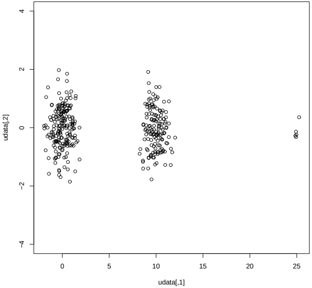
Therefore, outlier changes cov-matrix model.

5. Tuning constants

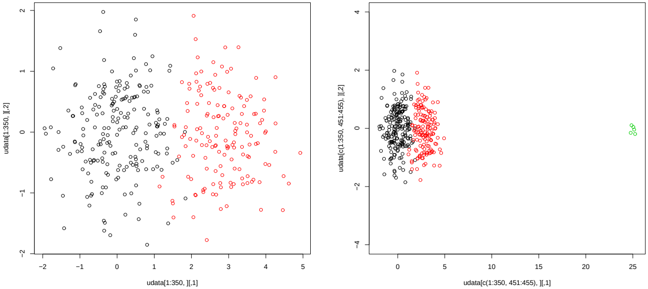
- Flexible cov-matrices: need lower bound for cov-matrix eigenvalue.
- Trimming methods need trimming rate.
- Improper noise method needs noise level c.

These tune robustness behaviour as well.

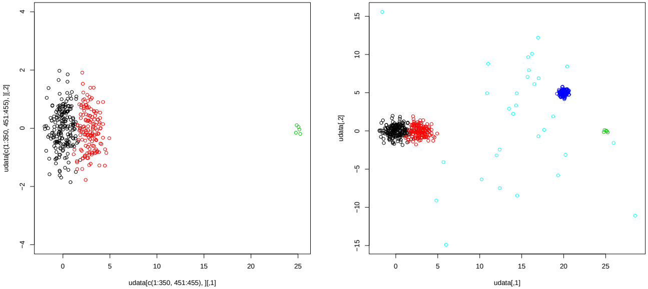




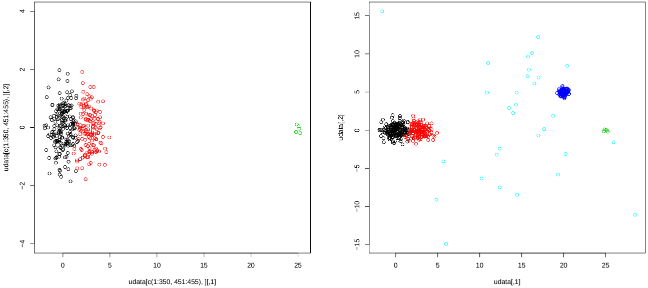
How many points are needed to form a "cluster", not a "group of outliers"?
(May depend on how packed they are.)



Rather separate non-well separated clusters?
(May depend on what's added.)



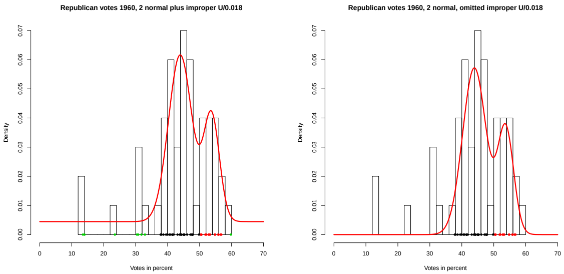
Large-variance: cluster or outliers?



Large-variance: cluster or outliers?

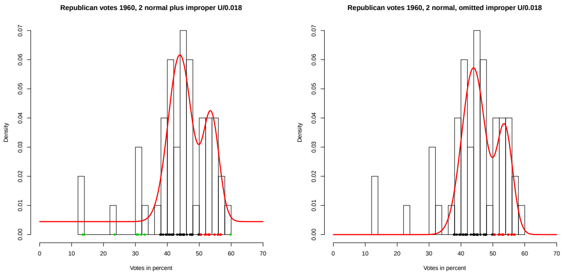
No proper robustness assessment without decision about what we want a method to do.
In CA we don't *want* 0.5 or 0.3 breakdown point.

Current work on improper noise (Coretto & Hennig)



Idea: fit noise so that remaining points are "good mixture" (see also Gallegos & Ritter 2005; χ^2/Mahalanobis)

Current work on improper noise (Coretto & Hennig)



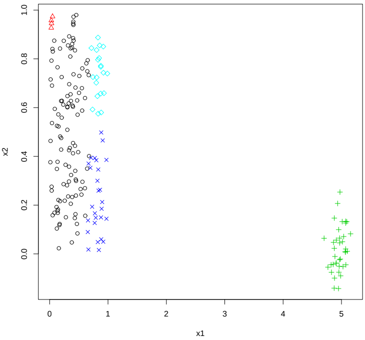
Idea: fit noise so that remaining points are "good mixture" (see also Gallegos & Ritter 2005; χ^2/Mahalanobis)

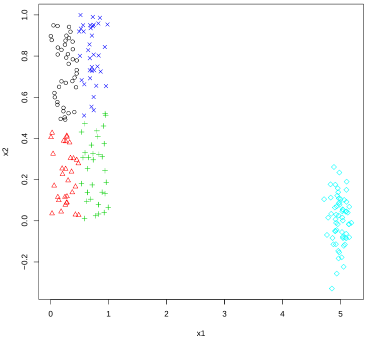
Problem: Data-dependent method, needs new robustness theory.

Interaction with estimating *s*?

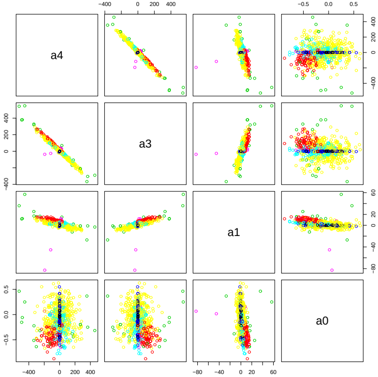
6. Conclusion

There are more nasty problems in robust clustering than just outliers.





Real data are more nasty than our example models.



Would like to have:

- Noise-mixture/trimming method with flexible cov-matrices and estimation of trimming level and s.

Would like to have:

- Noise-mixture/trimming method with flexible cov-matrices and estimation of trimming level and s.
- Such a method *will still need a tuning constant* governing “small cluster”/”group of outliers” trade-off.

Would like to have:

- Noise-mixture/trimming method with flexible cov-matrices and estimation of trimming level and s.
- Such a method *will still need a tuning constant* governing "small cluster"/"group of outliers" trade-off.
- Robustness results that show that this is better than plain Gaussian mixtures with estimated s.

Would like to have:

- Noise-mixture/trimming method with flexible cov-matrices and estimation of trimming level and s.
- Such a method *will still need a tuning constant* governing "small cluster"/"group of outliers" trade-off.
- Robustness results that show that this is better than plain Gaussian mixtures with estimated s.
- Stronger results to assess non-outlier-related (and non-worst case) instability/robustness

Would like to have:

- Noise-mixture/trimming method with flexible cov-matrices and estimation of trimming level and s.
- Such a method *will still need a tuning constant* governing “small cluster”/”group of outliers” trade-off.
- Robustness results that show that this is better than plain Gaussian mixtures with estimated s.
- Stronger results to assess non-outlier-related (and non-worst case) instability/robustness
- Always need decision what kind of clusters we want.