Benford's Law Conference

10-12 July 2019 - Stresa, Italy

Denoising and Trimming for Improved Cluster Solutions with Applications to Customs Frauds

Andrea Cerioli[1], Luis Ángel García-Escudero[2], Alfonso Gordaliza[2], Carlos Matrán[2], **Agustín Mayo-Iscar[2]**, Domenico Perrotta[3], Marco Riani[1] and Francesca Torti[3]

1. Department of Economics & Ro.S.A. University of Parma
2. Department of Statistics and O.R. & IMUVA. University of Valladolid
3. JRC European Comission. Ispra

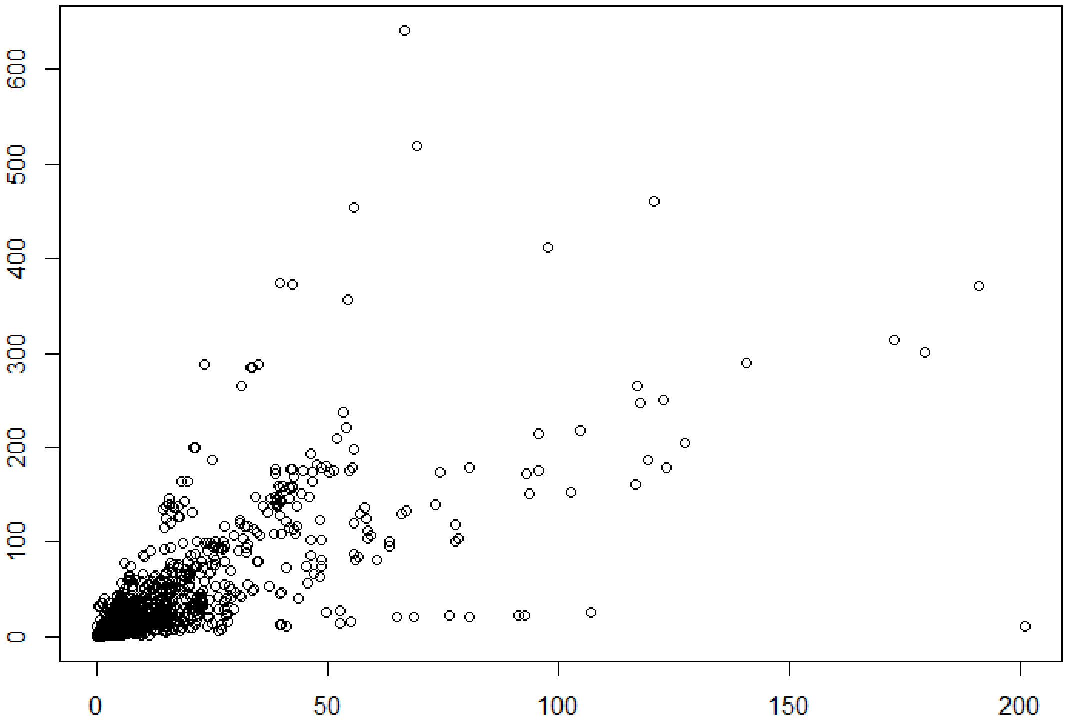
Plan of the talk

Trimmed & Constrained Maximum Likelihood (ML) proposals for Model Based Clustering

Robustness is based on the joint application of trimming & constraints

Clustering of Regression models

Application to Comext data

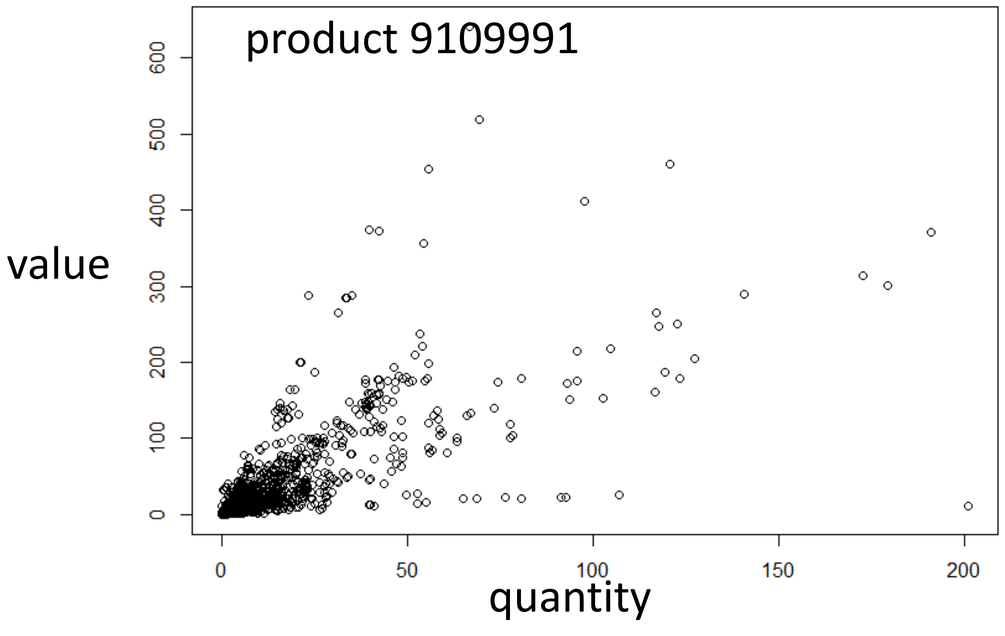


ComExt database

The ComExt Extra-European trade database provides statistics on merchandise trade among European Union member states, and between member states and global partners. ComExt, published by Eurostat, is based on data provided by the statistical agencies of the EU member states and trading partners. The statistics of interest for anti-fraud are mainly the traded **volumes and values for a fixed product**, which are aggregated monthly by Eurostat.

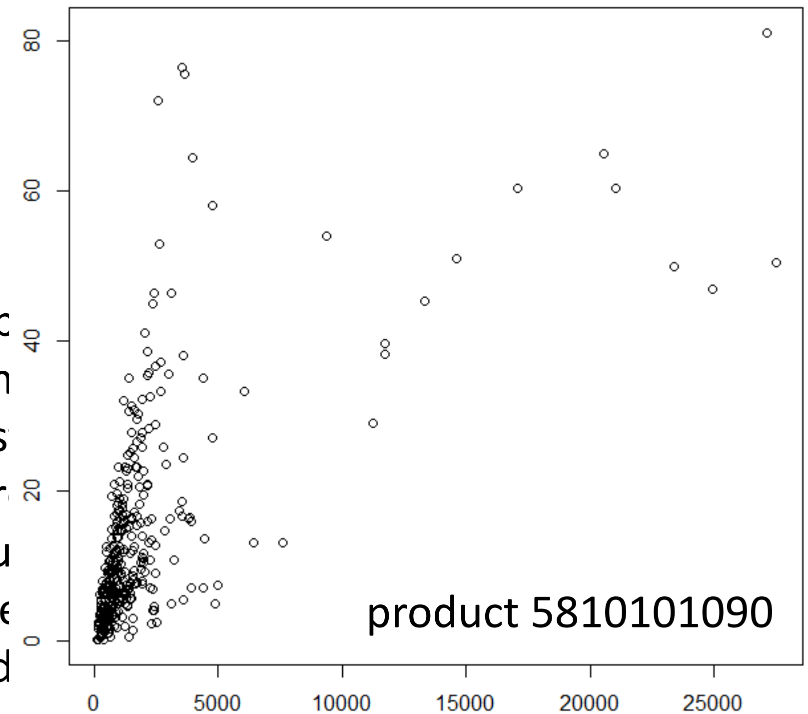
We are interested in applying robust clustering procedures for identifying outliers in the ComExt Extra-European trade database by thinking about its usefulness in fraud detection.

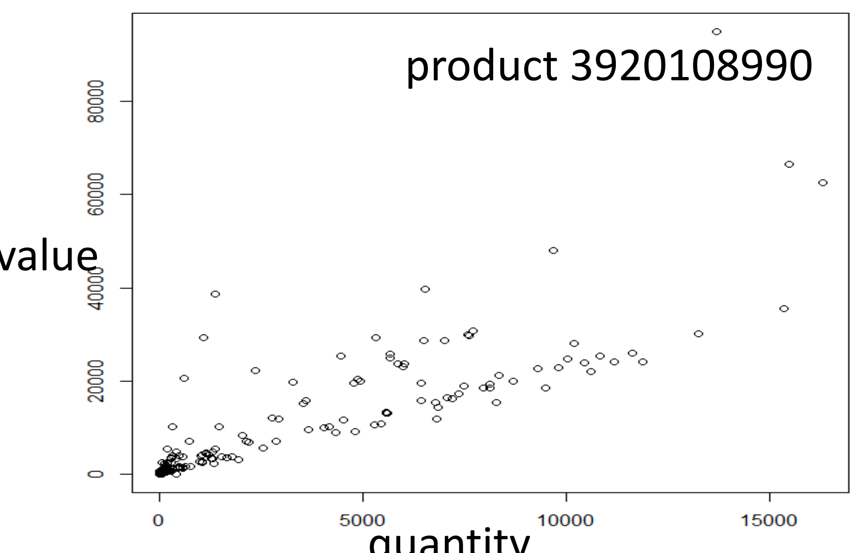
There are robust procedures available for clustering data in different settings, including ones devoted to identifying clusters around linear subspaces which appear to be well suited for datasets in this database.

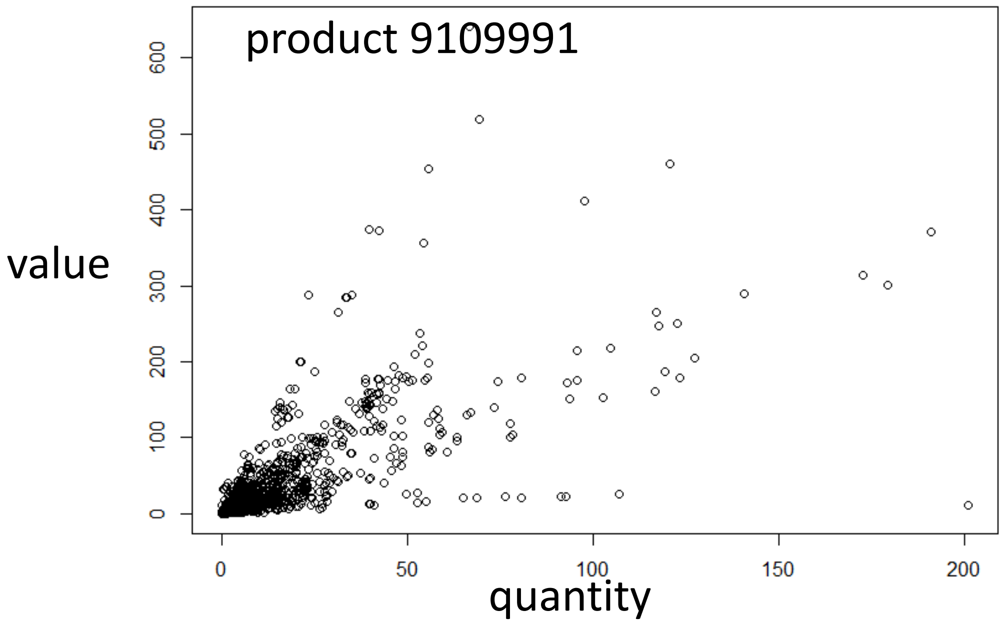


ComExt database

The ComExt Extra-European trade database provides European Union member states, and between memb published by Eurostat, is based on data provided by th states and trading partners. The statistics of interes **volumes and values for a fixed product**, which are aggr We are interested in applying robust clustering procedu Extra-European trade database by thinking about its use There are robust procedures available for clustering d devoted to identifying clusters around linear subspaces which appear to be well suited for datasets in this database.

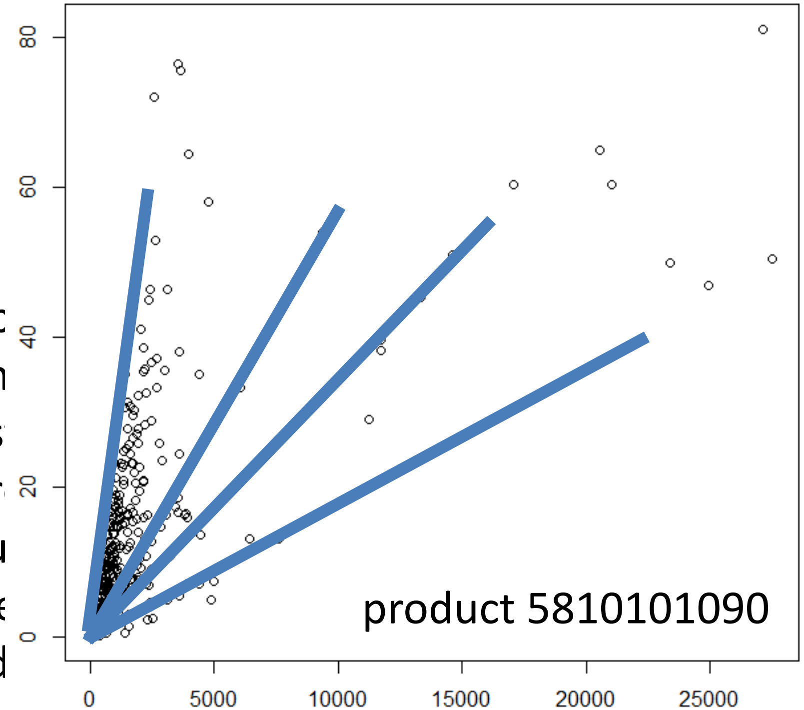


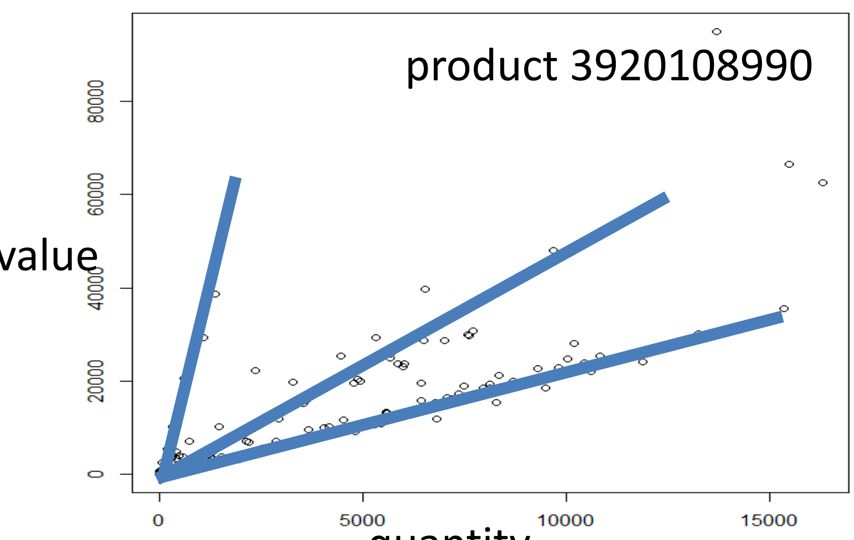


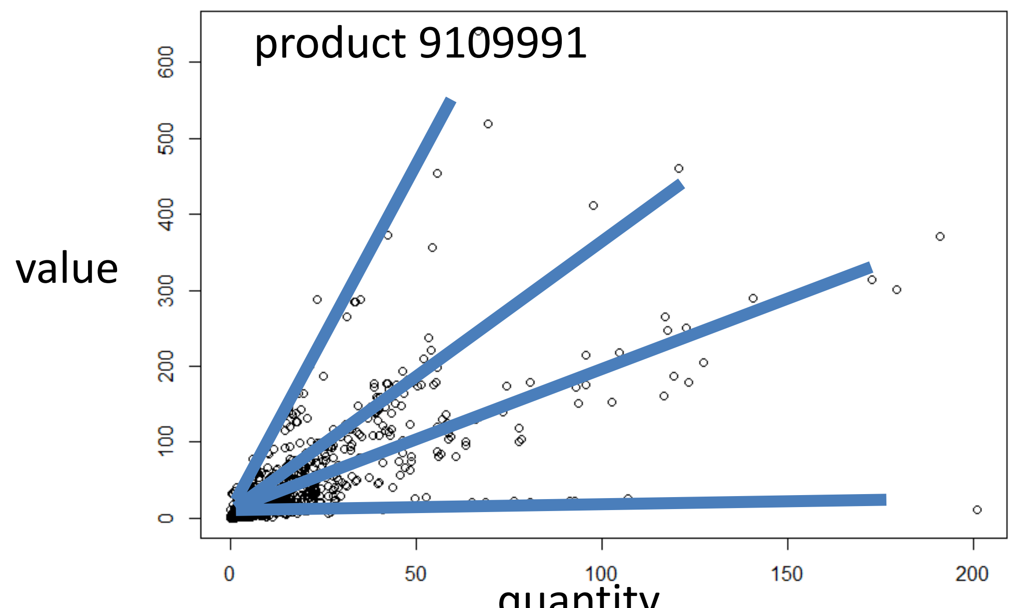


ComExt database

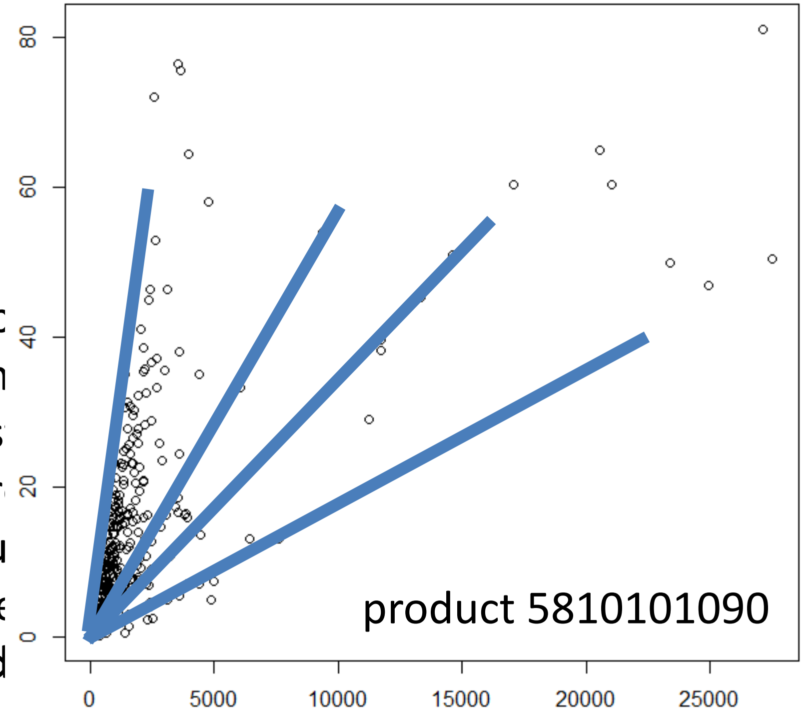
The ComExt Extra-European trade database provides
European Union member states, and between memb
published by Eurostat, is based on data provided by th
states and trading partners. The statistics of interes
volumes and values for a fixed product, which are aggr
We are interested in applying robust clustering procedu
Extra-European trade database by thinking about its use
There are robust procedures available for clustering d
devoted to identifying clusters around linear subspaces which appear to be well suited for
datasets in this database.



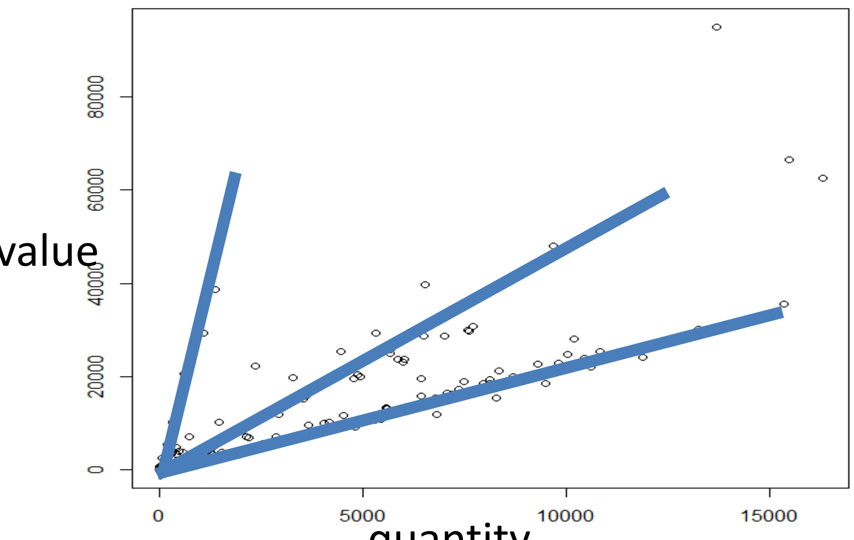


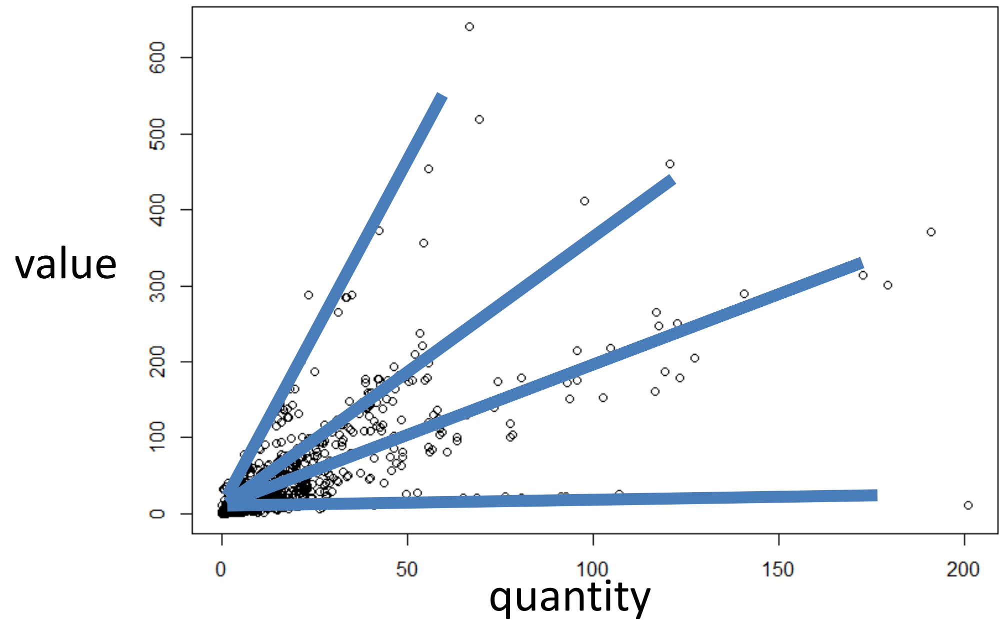


ComExt database



The ComExt Extra-European trade database provides
European Union member states, and between memb
published by Eurostat, is based on data provided by th
states and trading partners. The statistics of interes
volumes and values for a fixed product, which are aggr
We are interested in applying robust clustering procedu
Extra-European trade database by thinking about its use
There are robust procedures available for clustering d
devoted to identifying clusters around linear subspaces which appear to be well suited for datasets in this database.

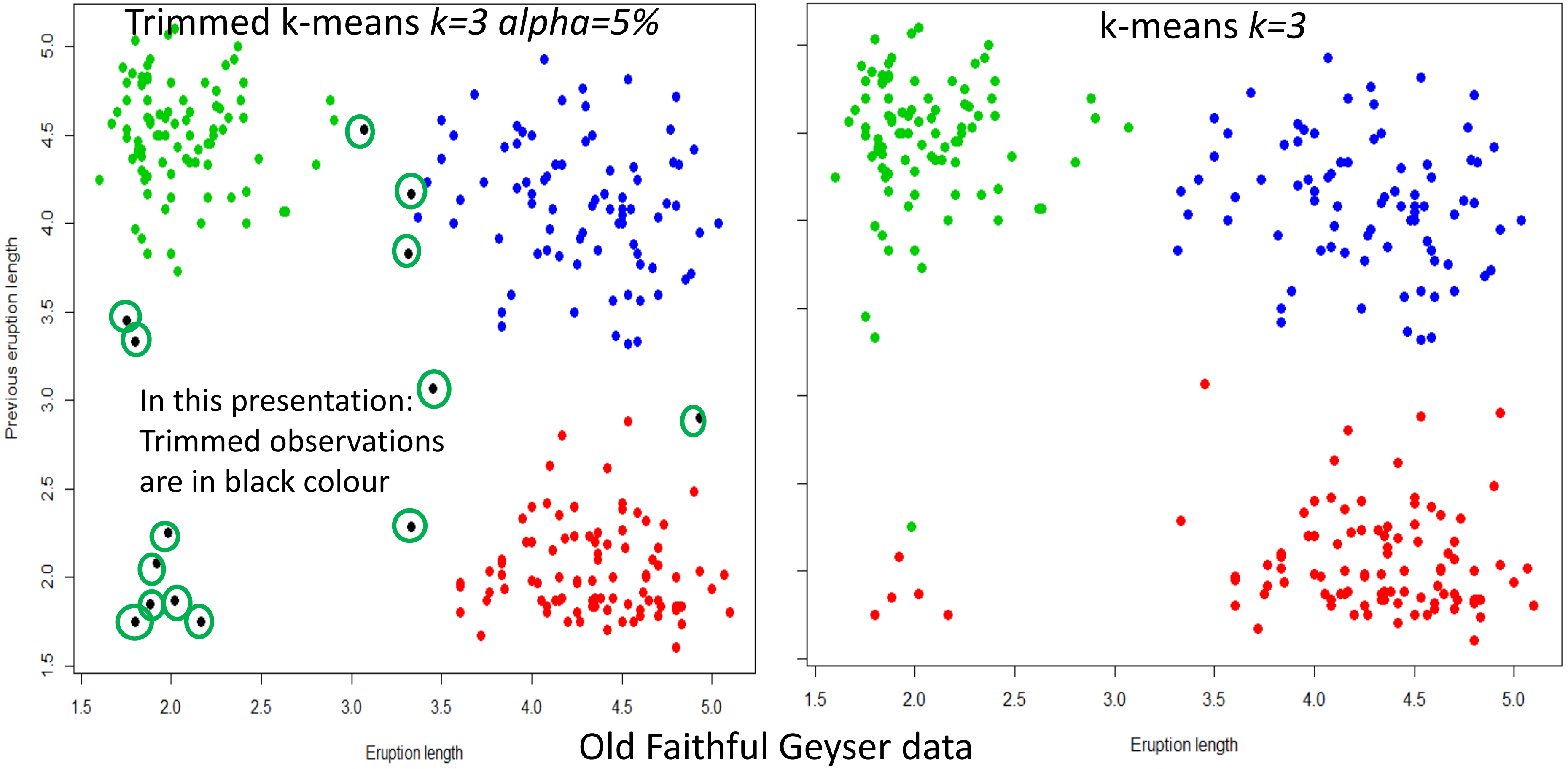




Trimmed k-means

Trimmed k-means. Cuesta, Gordaliza and Matran (1997) Trimmed k-means: an attempt to robustify quantizers. The Annals of Statistics, 25(2), 553-576.

A robust release of k-means

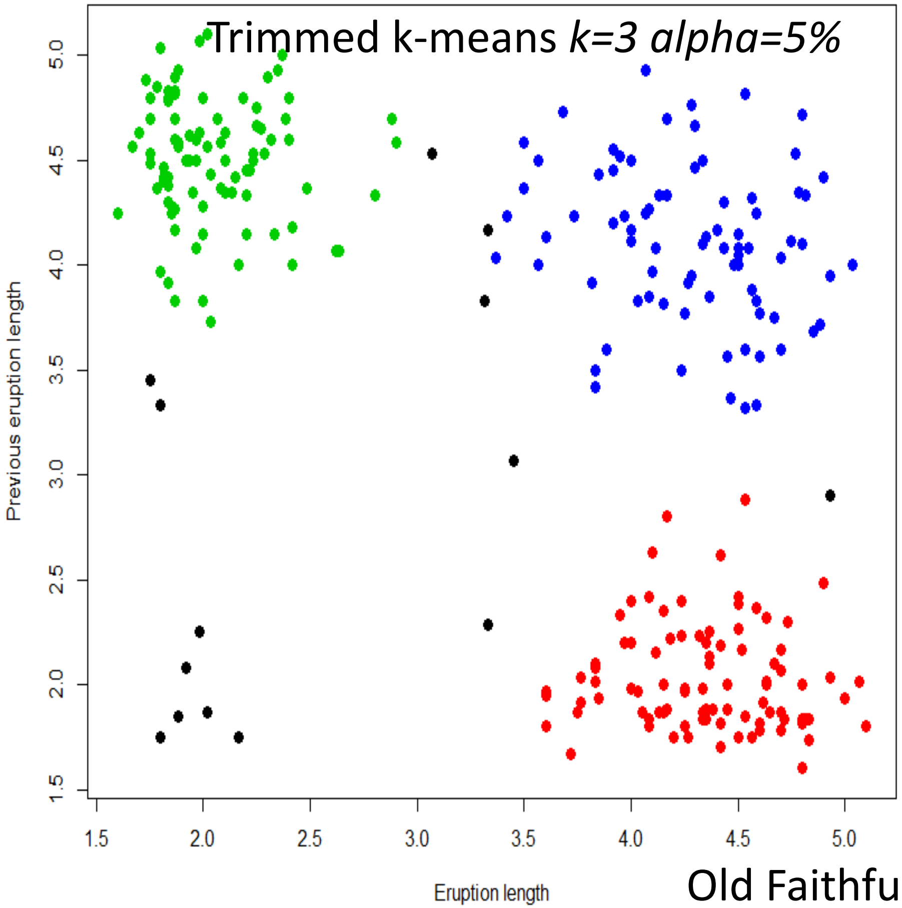


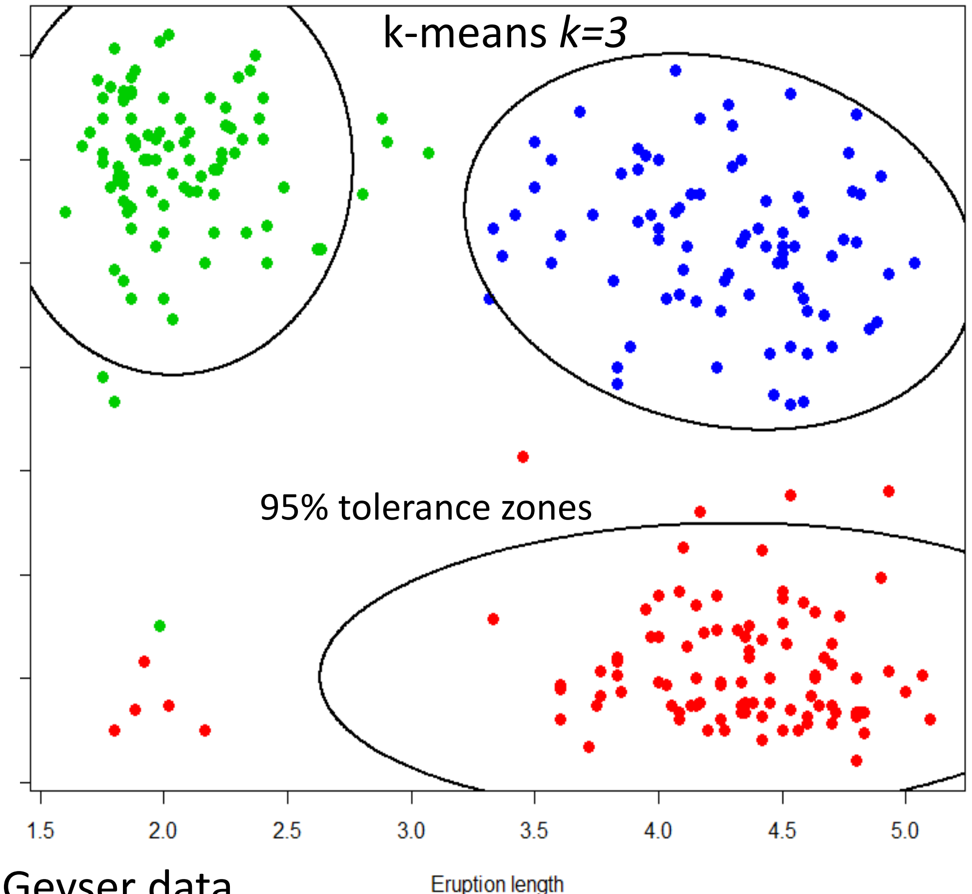
Old Faithful Geyser data

Trimmed k-means

Trimmed k-means. Cuesta, Gordaliza and Matran (1997) Trimmed k-means: an attempt to robustify quantizers. The Annals of Statistics, 25(2), 553-576.

A robust release of k-means



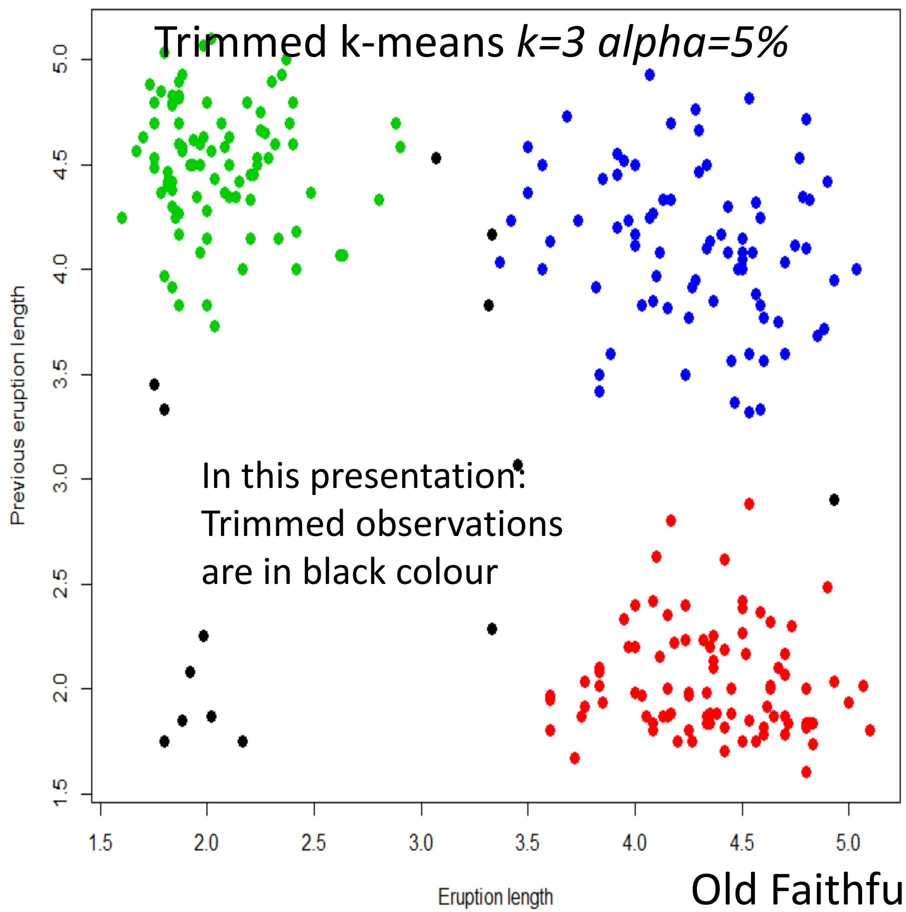


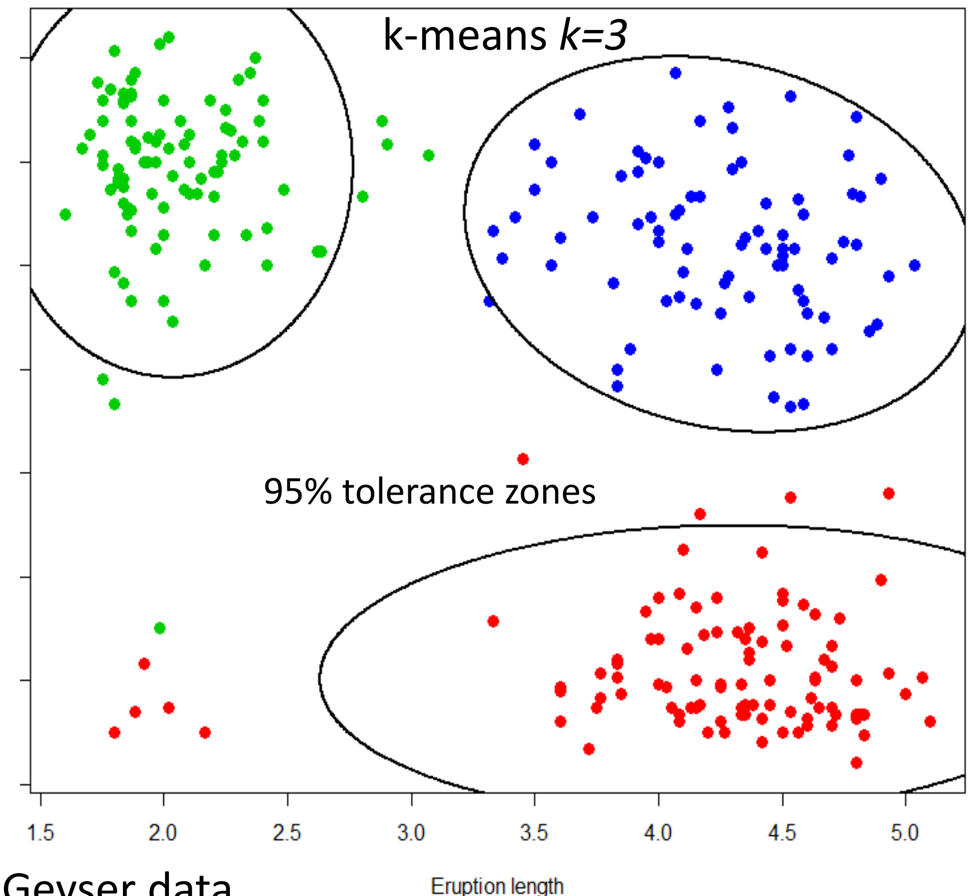
Old Faithful Geyser data

Trimmed k-means

Trimmed k-means. Cuesta, Gordaliza and Matran (1997) Trimmed k-means: an attempt to robustify quantizers. The Annals of Statistics, 25(2), 553-576.

A robust release of k-means



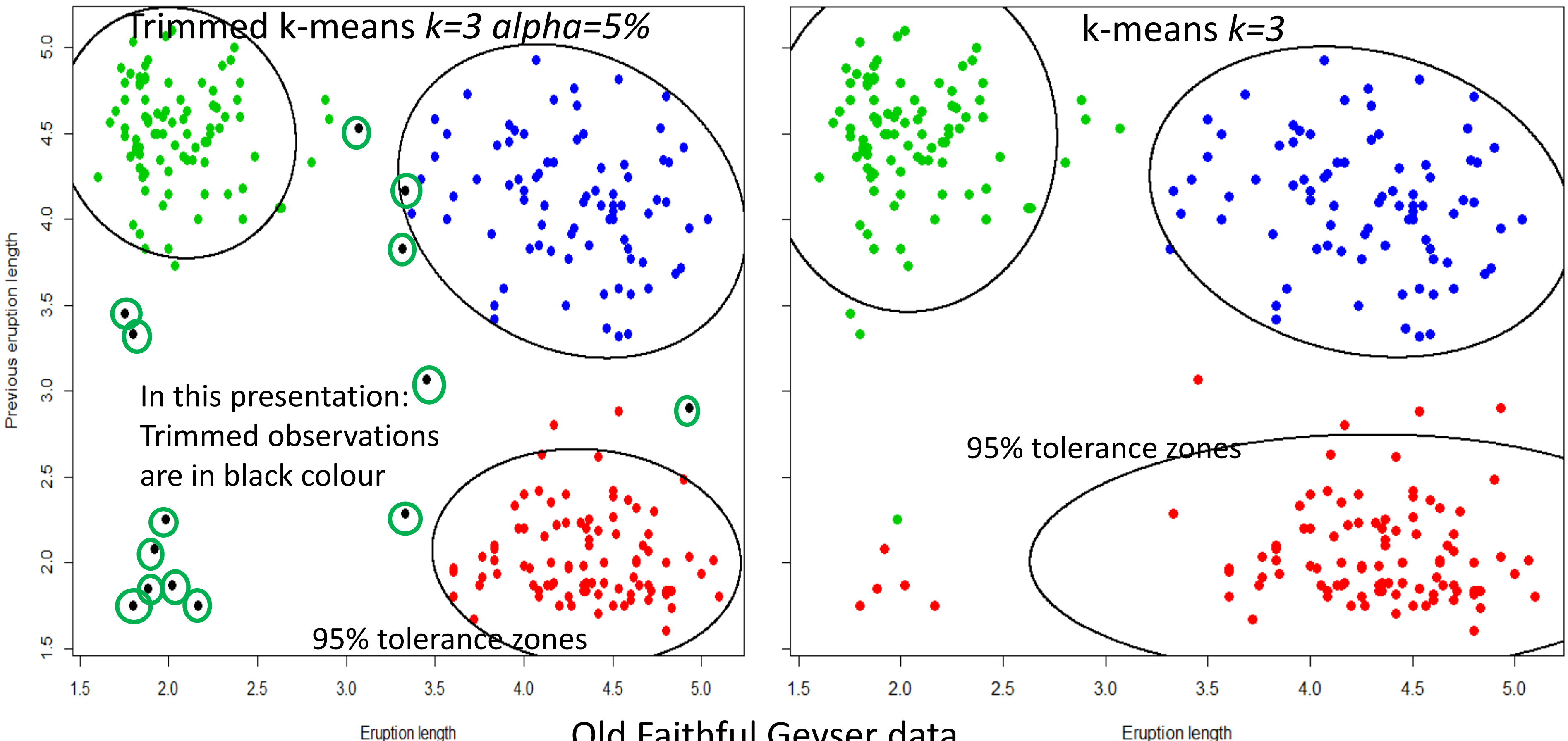


Old Faithful Geyser data

Trimmed k-means

Trimmed k-means. Cuesta, Gordaliza and Matran (1997) Trimmed k-means: an attempt to robustify quantizers. The Annals of Statistics, 25(2), 553-576.

A robust release of k-means



Old Faithful Geyser data

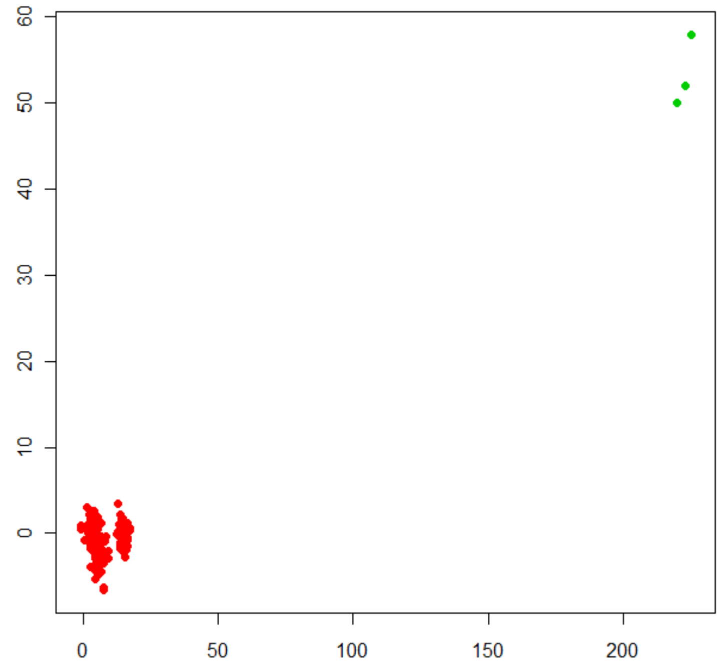
Trimmed k-means

k-means is not a robust procedure.

Possible effects produced by contamination in clustering:

- To change the estimation of location and scatter corresponding to each true group
- To merge several real groups in one component of the solution

Alternatives based on M-estimators increase the resistance against the influence of outliers.

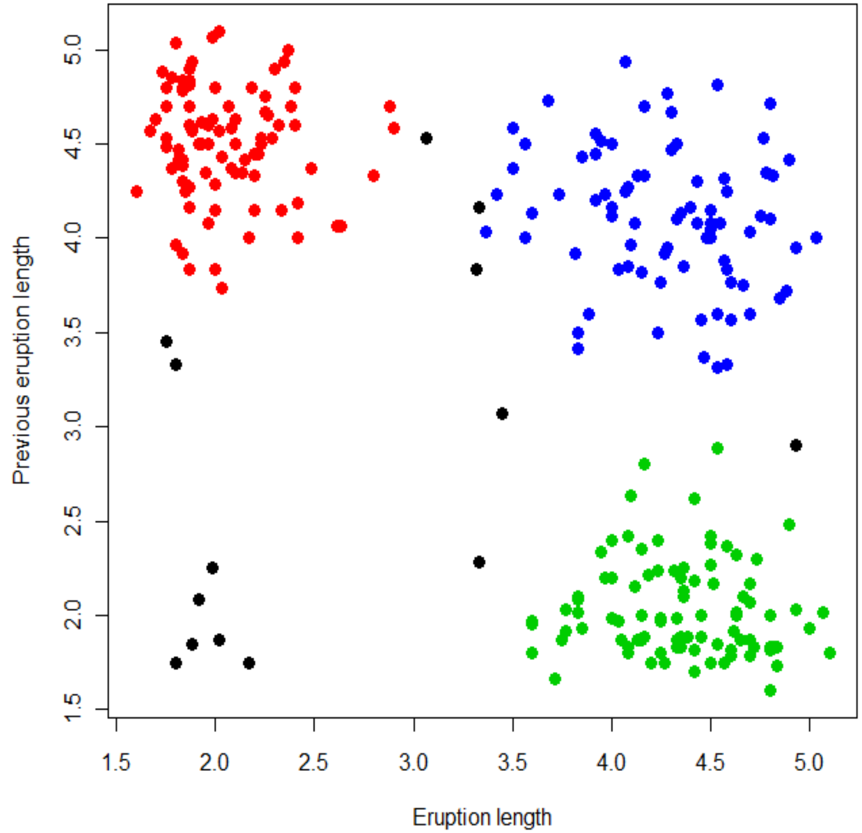


PAM *k=2*
Partitioning Around Medoids

But, our recommendation is to use trimming for avoiding the influence of contamination in the cluster parameters estimation. The level of trimming, α, is given in advance and has to be greater than the contamination level.

Trimmed k-means

Robustness is based on impartial trimming: the sample decides which is the best way to trim.



Trimmed k-means *k=3 alpha=5%*

Given a sample $\{x_1 \ldots x_i \ldots x_n\}$

to find the best *k* quantizers

in the sense of $(\mu_1 \ldots \mu_j \ldots \mu_k)$

$$\arg\min_{\mu, z} \min_{A / P_n(A) = 1-\alpha} \sum_{i=1}^{n} I_A(x_i) \sum_{j=1}^{k} z_{ij} \left\| x_i - \mu_j \right\|^2$$

where

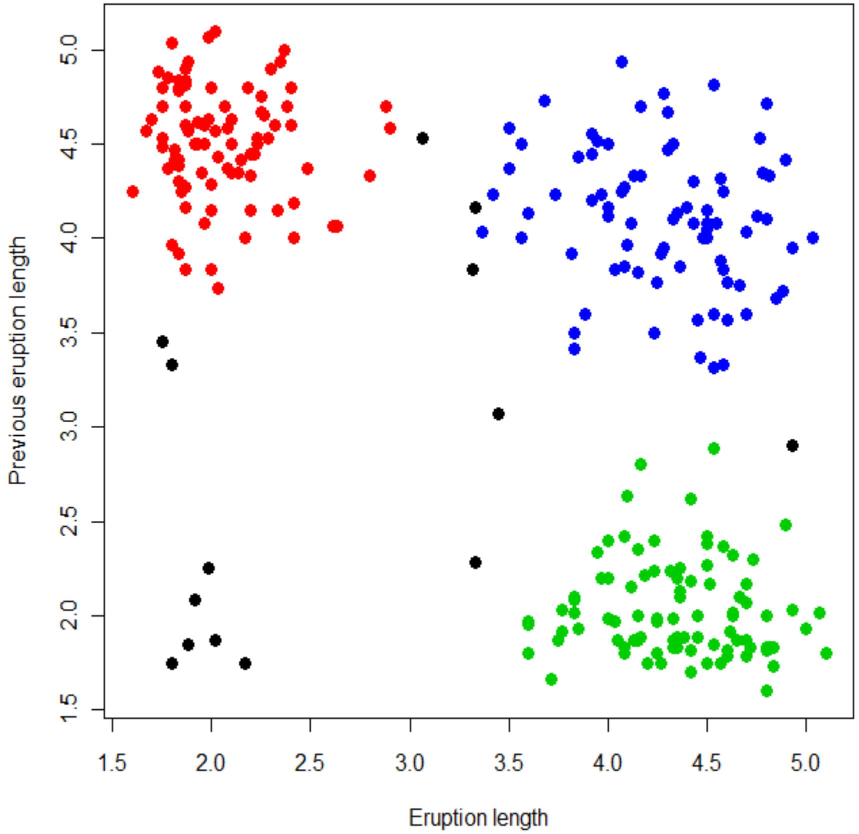
$z_{ij} = 0 \; or \; 1$ with $\sum_{j=1}^{k} z_{ij} = 1$ defines the assignment $I_A(x)$ Indicator of belonging to *A*

A is a set with size 1-α containing the non-trimmed observations.

Double search: the best way of trimming and the best quantizers and assignment

Trimmed k-means

Robustness is based on impartial trimming: the sample decides which is the best way to trim.



Trimmed k-means *k=3 alpha=5%*

Given a sample $\{x_1 \ldots x_i \ldots x_n\}$

to find the best *k* quantizers

in the sense of $(\mu_1 \ldots \mu_j \ldots \mu_k)$

$$\arg\min_{\mu, z} \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \left\| x_i - \mu_j \right\|^2$$

where

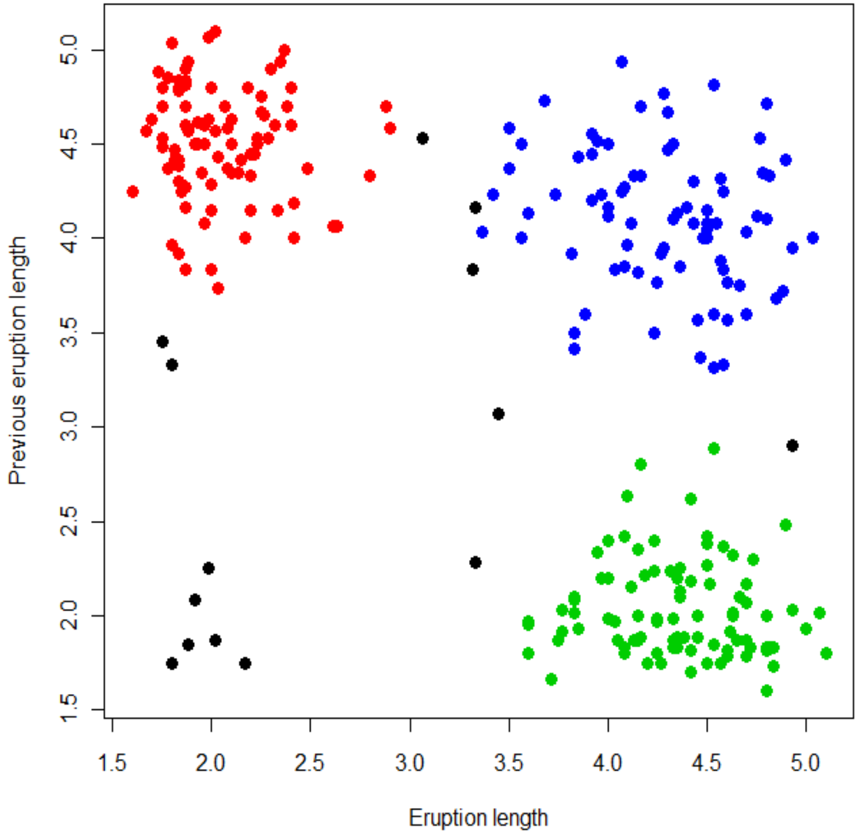
$z_{ij} = 0 \; or \; 1$ with $\sum_{j=1}^{k} z_{ij} = 1$ defines the assignment $I_A(x)$ Indicator of belonging to *A*

A is a set with size 1-α containing the non-trimmed observations.

Double search: the best way of trimming and the best quantizers and assignment

Trimmed k-means

Robustness is based on impartial trimming: the sample decides which is the best way to trim.



Trimmed k-means *k=3 alpha=5%*

Given a sample $\{x_1 \ldots x_i \ldots x_n\}$

to find the best *k* quantizers

in the sense of $(\mu_1 \ldots \mu_j \ldots \mu_k)$

$$\arg\min_{\mu, z} \min_{A / P_n(A) = 1-\alpha} \sum_{i=1}^{n} I_A(x_i) \sum_{j=1}^{k} z_{ij} \left\| x_i - \mu_j \right\|^2$$

where

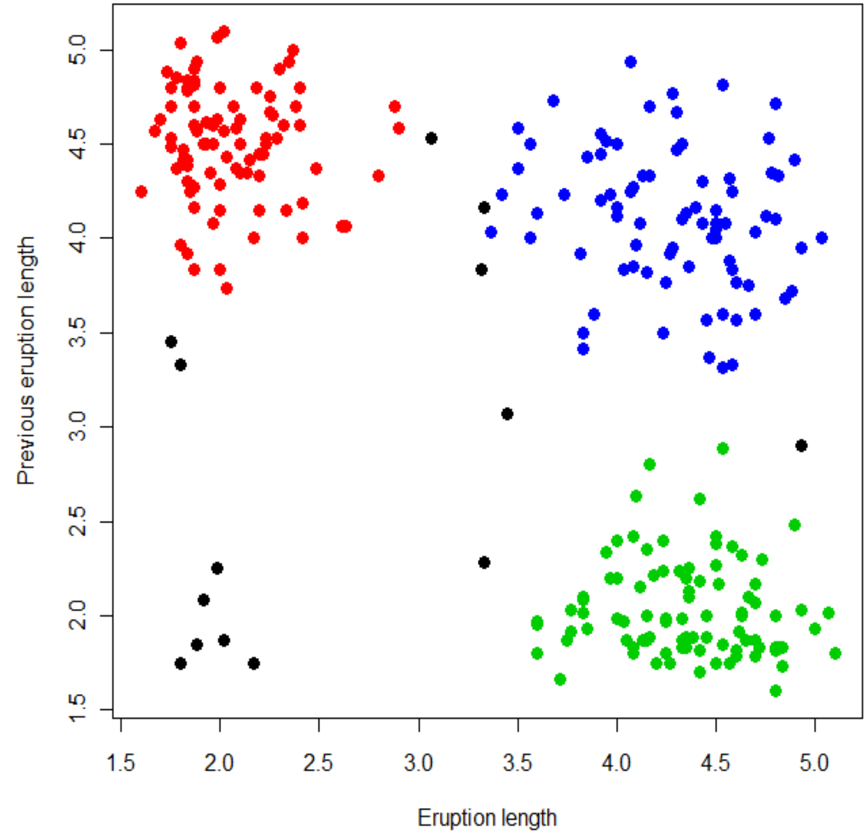
$z_{ij} = 0 \; or \; 1$ with $\sum_{j=1}^{k} z_{ij} = 1$ defines the assignment $I_A(x)$ Indicator of belonging to *A*

A is a set with size 1-α containing the non-trimmed observations.

Double search: the best way of trimming and the best quantizers and assignment

Trimmed k-means

Robustness is based on impartial trimming: the sample decides which is the best way to trim.



Trimmed k-means *k=3 alpha=5%*

Given a sample $\{x_1 \ldots x_i \ldots x_n\}$

to find the best *k* quantizers

in the sense of $(\mu_1 \ldots \mu_j \ldots \mu_k)$

$$\arg\min_{\mu, z} \min_{A / P_n(A) = 1-\alpha} \sum_{i=1}^{n} I_A(x_i) \sum_{j=1}^{k} z_{ij} \left\| x_i - \mu_j \right\|^2$$

Impartial trimming

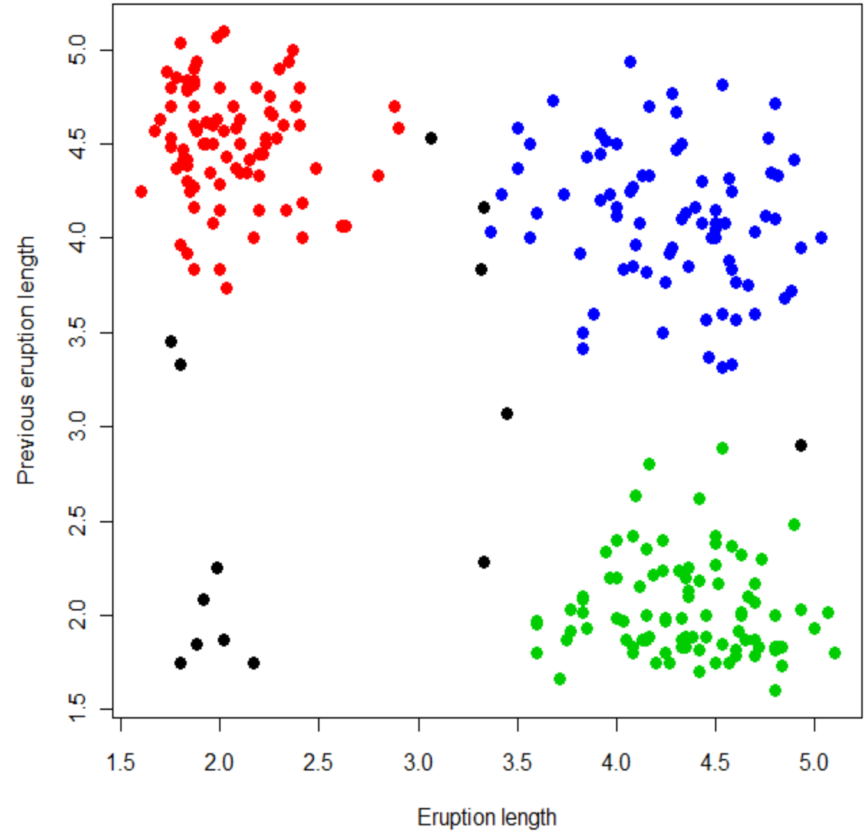
where

$z_{ij} = 0 \; or \; 1$ with $\sum_{j=1}^{k} z_{ij} = 1$ defines the assignment $I_A(x)$ Indicator of belonging to *A*

A is a set with size 1-α containing the non-trimmed observations.

Double search: the best way of trimming and the best quantizers and assignment

Trimmed k-means



Robustness is based on impartial trimming: the sample decides which is the best way to trim.

Given a sample $\{x_1 \ldots x_i \ldots x_n\}$

To find the best k quantizers

in the sense of $(\mu_1 \ldots \mu_j \ldots \mu_k)$

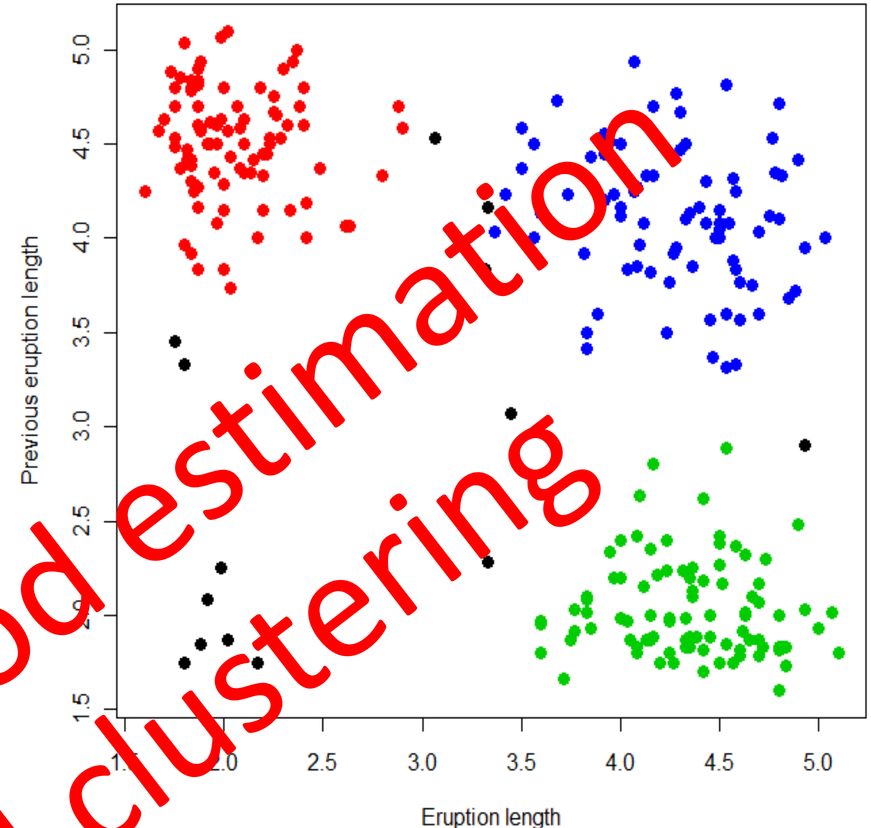
$$\arg\min_{\mu, z} \min_{A / P_n(A) = 1-\alpha} \sum_{i=1}^{n} I_A(x_i) \sum_{j=1}^{k} z_{ij} \left\| x_i - \mu_j \right\|^2$$

$$\arg\sup_{\mu, z} \sup_{A / P_n(A) = 1-\alpha} \sum_{i=1}^{n} I_A(x_i) \sum_{j=1}^{k} z_{ij} \log\left(\pi_j N_{\mu_j, \Sigma_j}(x_i) \right)$$

$N_{\mu,\Sigma}$ Normal density

with the constraints $\Sigma_1 = .. = \Sigma_j = .. = \Sigma_k = \lambda I_p$, $\pi_1 = \ldots = \pi_j = \ldots = \pi_k$

Trimmed k-means



Robustness is based on impartial trimming: the sample decides which is the best way to trim.

Given a sample $\{x_1 \ldots x_i \ldots x_n\}$

To find the best k quantizers

in the sense of $(\mu_1 \ldots \mu_j \ldots \mu_k)$

$$\arg\min_{\mu,z} \min_{A/P_n(A)=1-\alpha} \sum_{i=1}^{n} I_A(x_i) \sum_{j=1}^{k} z_{ij} \left\| x_i - \mu_j \right\|^2$$

$$\arg\sup_{\mu,z} \sup_{A/P_n(A)=1-\alpha} \sum_{i=1}^{n} I_A(x_i) \sum_{j=1}^{k} z_{ij} \log\left(\pi_j N_{\mu_j,\Sigma_j}(x_i)\right)$$

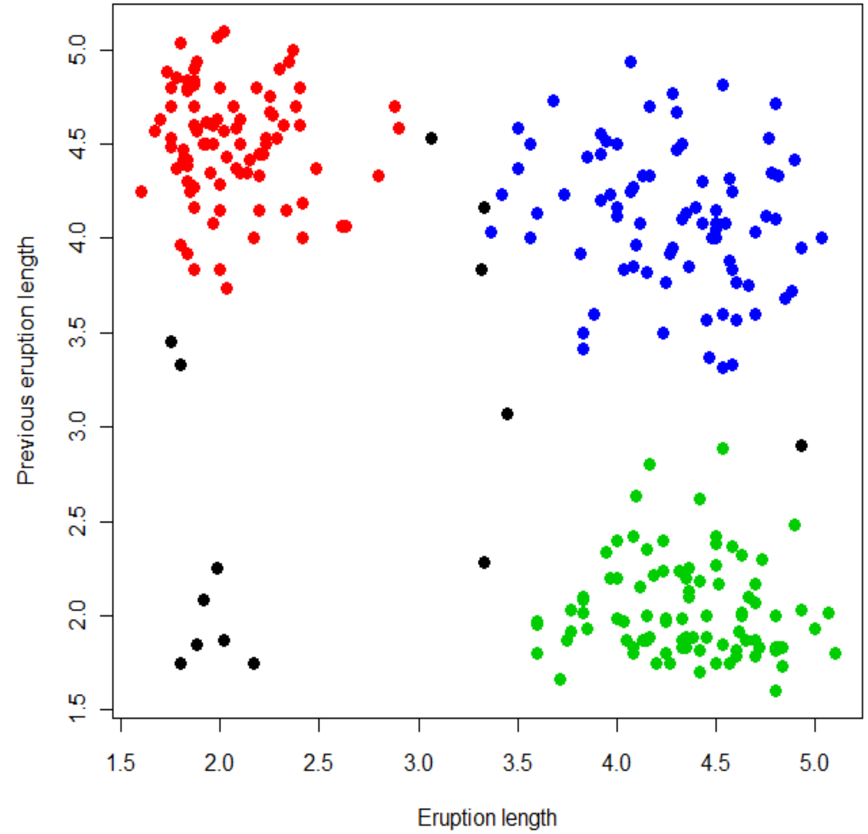
$N_{\mu,\Sigma}$ Normal density

with the constraints $\Sigma_1 = .. = \Sigma_j = .. = \Sigma_k = \lambda I_p$, $\pi_1 = \ldots = \pi_j = \ldots = \pi_k$

Maximum likelihood estimation

Model based clustering

Trimmed k-means



Robustness is based on impartial trimming: the sample decides which is the best way to trim.

Given a sample $\{x_1 \ldots x_i \ldots x_n\}$

To find the best *k* quantizers

in the sense of $(\mu_1 \ldots \mu_j \ldots \mu_k)$

$$\arg\min_{\mu,z} \min_{A/P_n(A)=1-\alpha} \sum_{i=1}^{n} I_A(x_i) \sum_{j=1}^{k} z_{ij} \left\| x_i - \mu_j \right\|^2$$

Trimming & constraints

$$\arg\sup_{\mu,z} \sup_{A/P_n(A)=1-\alpha} \sum_{i=1}^{n} I_A(x_i) \sum_{j=1}^{k} z_{ij} \log\left(\pi_j N_{\mu_j,\Sigma_j}(x_i)\right)$$

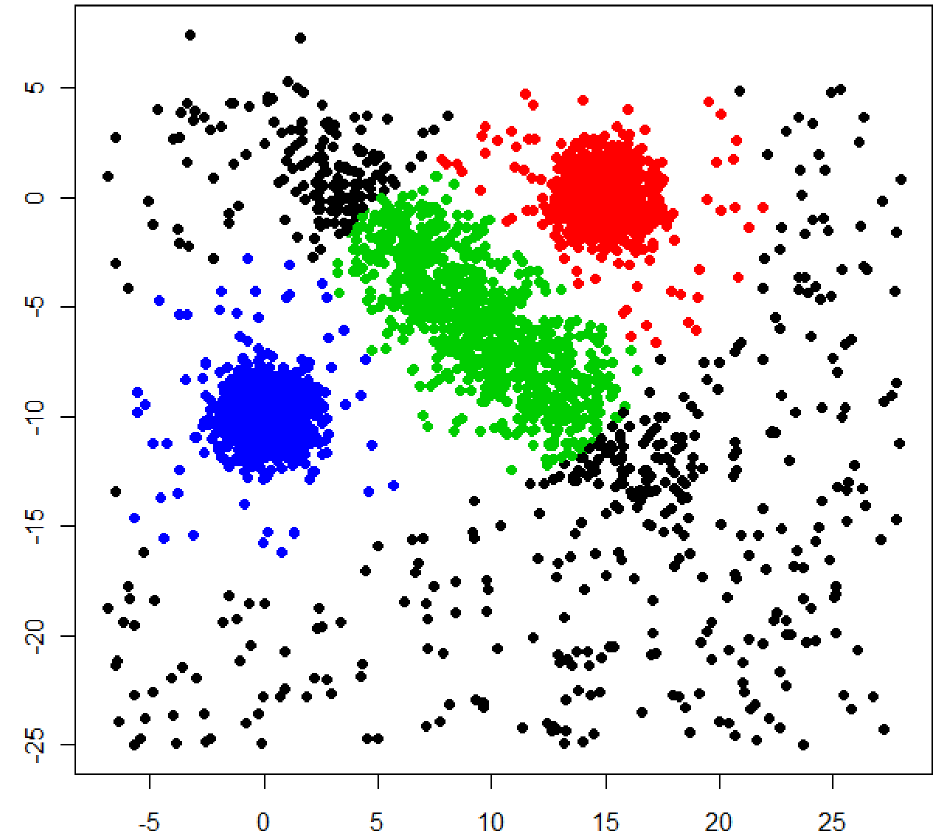
with the constraints $\Sigma_1 = .. = \Sigma_j = .. = \Sigma_k = \lambda I_p$

$\pi_1 = \ldots = \pi_j = \ldots = \pi_k$

Very strong constraints

Trimmed k-means

Robustness is based on impartial trimming: the sample decides which is the best way to trim.



Trimmed k-means *k=3 alpha=15%*

Given a sample $\{x_1 \ldots x_i \ldots x_n\}$

To find the best *k* quantizers

in the sense of $(\mu_1 \ldots \mu_j \ldots \mu_k)$

$$\arg\min_{\mu,z} \min_{A/P_n(A)=1-\alpha} \sum_{i=1}^{n} I_A(x_i) \sum_{j=1}^{k} z_{ij} \left\| x_i - \mu_j \right\|^2$$

Trimming & constraints

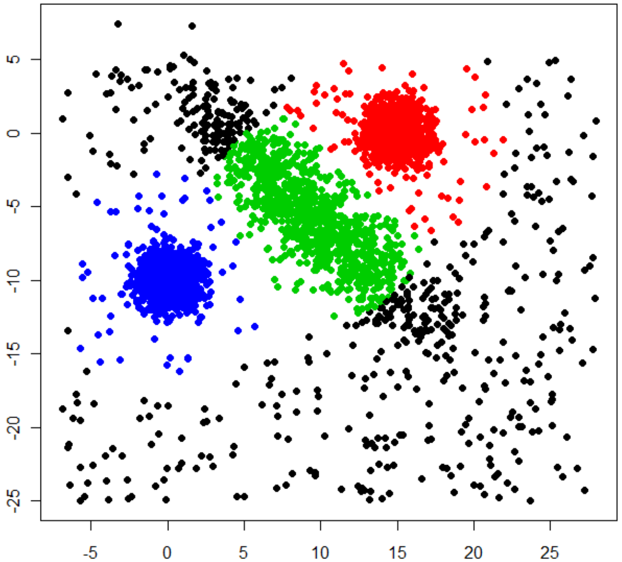
$$\arg\sup_{\mu,z} \sup_{A/P_n(A)=1-\alpha} \sum_{i=1}^{n} I_A(x_i) \sum_{j=1}^{k} z_{ij} \log\left(\pi_j N_{\mu_j,\Sigma_j}(x_i)\right)$$

with the constraints $\Sigma_1 = .. = \Sigma_j = .. = \Sigma_k = \lambda I_p$

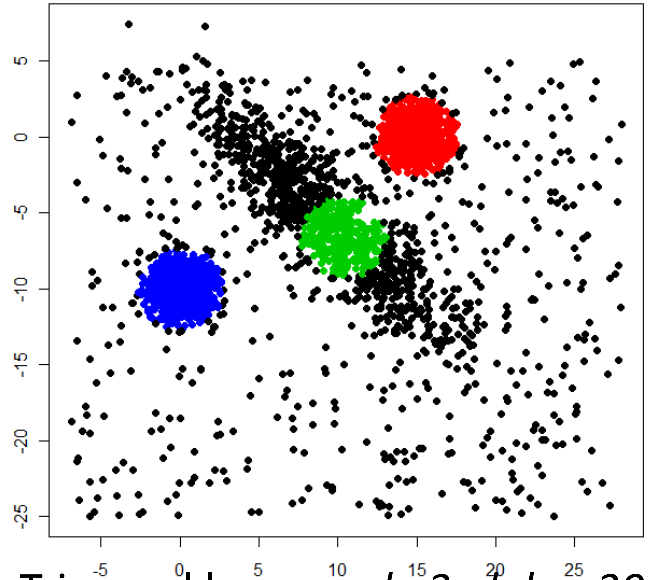
$\pi_1 = \ldots = \pi_j = \ldots = \pi_k$

Very strong constraints

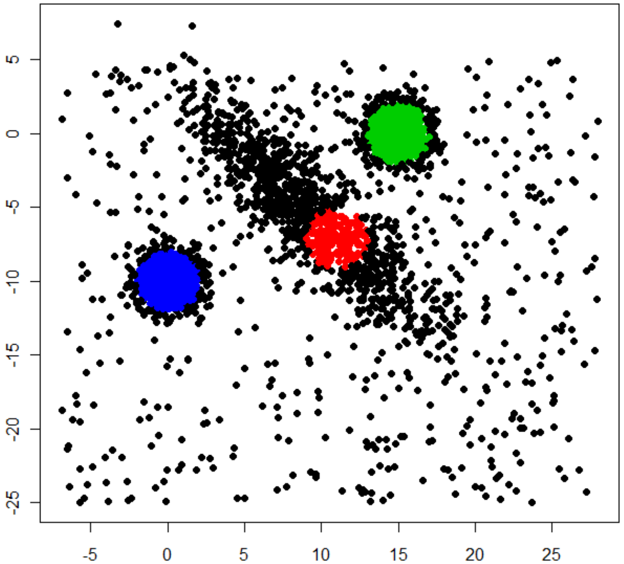
Trimmed k-means



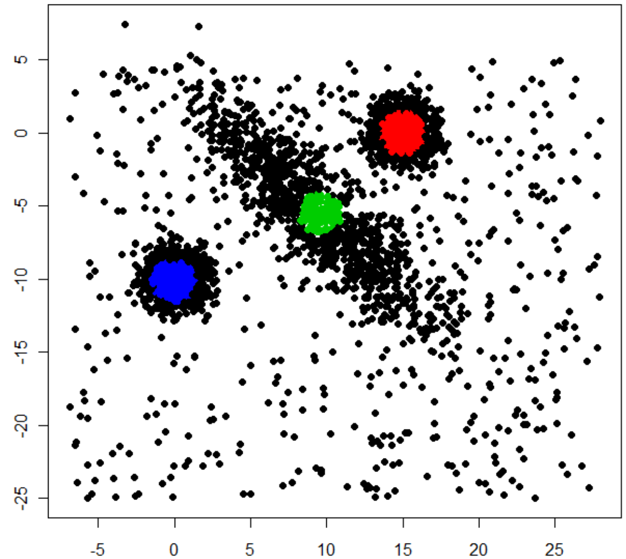
Trimmed k-means *k=3 alpha=15%*



Trimmed k-means *k=3 alpha=30%*



Trimmed k-means *k=3 alpha=45%*



Trimmed k-means *k=3 alpha=60%*

TCLUST methodology

García-Escudero, Gordaliza, Matrán and M-I (Annals of Stat. 2008).

- Estimator (likelihood based)

$$\arg\sup_{\theta\in R,\, z}\ \sup_{A/\#A=n(1-\alpha)} \sum_{i=1}^{n}\sum_{j=1}^{k} I_A(x_i) z_{ij} \log\left(\pi_j N^{p}_{\mu_j,\Sigma_j}(x_i)\right)$$

~~$\Sigma_1 = \ldots = \Sigma_j = \ldots = \Sigma_k = \lambda I_p$~~

~~$\pi_1 = \ldots = \pi_j = \ldots = \pi_k$~~

TCLUST: weaker constraints

where k is the number of components

$x_1, x_2, \ldots, x_n$ is a random sample

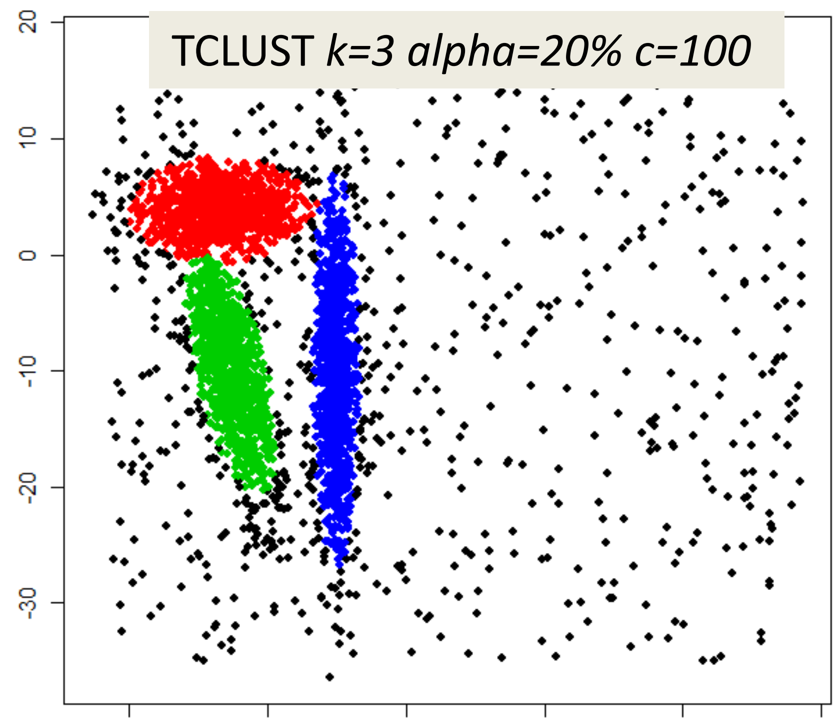
$N^{p}_{\mu,\Sigma}(x)$ is the Normal density

$$\theta = \left(\pi_1, \mu_1, \Sigma_1, \ldots, \pi_j, \mu_j, \Sigma_j, \ldots, \pi_k, \mu_k, \Sigma_k\right)$$

- Missing information
  - Membership z_{ij}

    which verifies $\sum_{j=1}^{k} z_{ij} = 1$ & $z_{ij} = 0 \text{ or } 1$
  - Genuineness $I_A(x_i)$

    which verifies $\sum_{i=1}^{n} I_A(x_i) = n(1-\alpha)$



Mixture of normal distributions

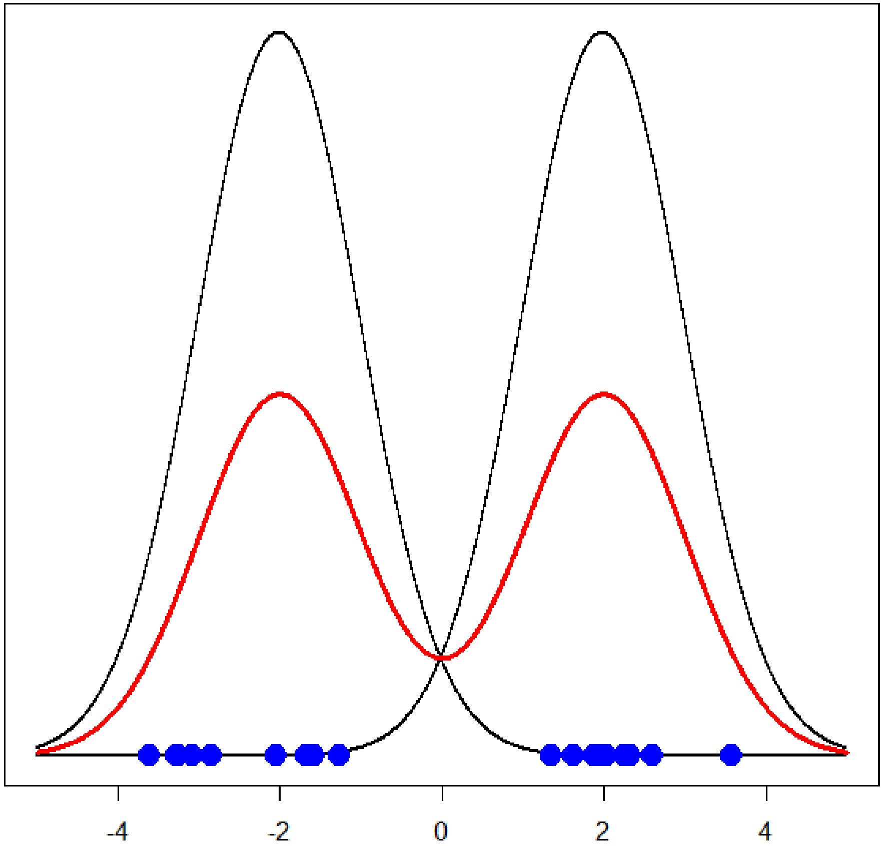
Density of a mixture of normal distributions

$$\sum_{j=1}^{G} \pi_j \, N_{\mu_j,\Sigma_j}(x)$$

$N_{\mu,\Sigma}(x)$ is the density of a Normal distribution with parameters μ and Σ

Mixture distribution parameter $\psi = \{\theta_1, \theta_2, \ldots, \theta_G\}$

where $\theta_j = (\pi_j, \mu_j, \Sigma_j) \quad j = 1..G$



density of a mixture of normals (red) and normal components (black)
points from a random sample of the mixture (blue)

Mixture of normal distributions

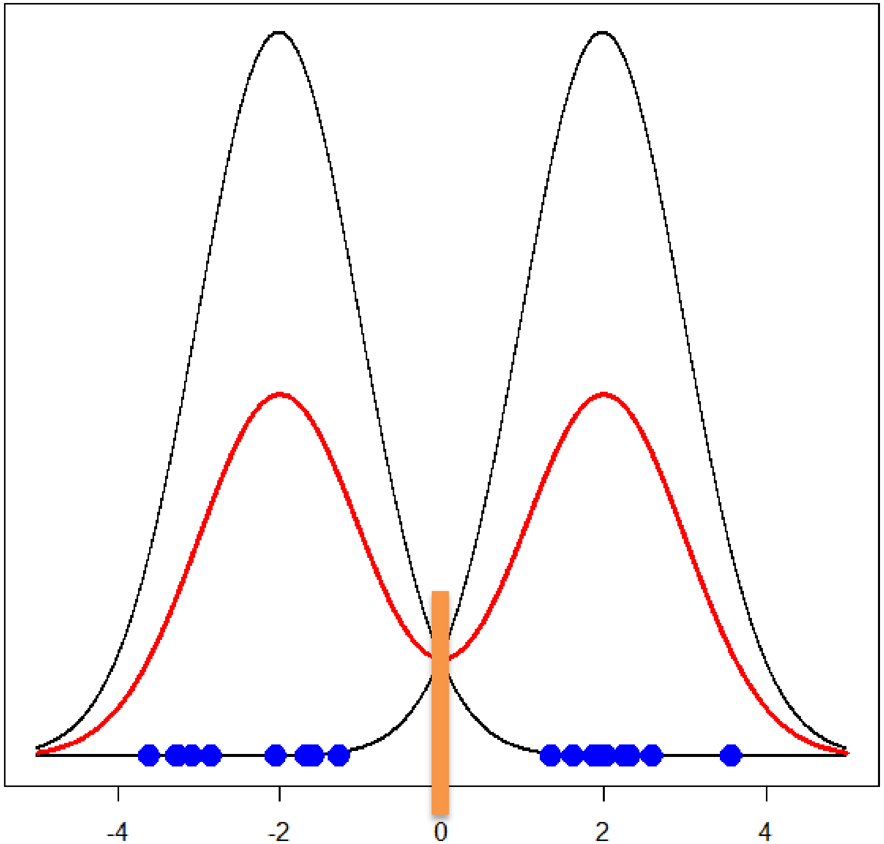
Density of a mixture of normal distributions

$$\sum_{j=1}^{G} \pi_j \, N_{\mu_j,\Sigma_j}(x)$$

$N_{\mu,\Sigma}(x)$ is the density of a Normal distribution with parameters μ and Σ

Mixture distribution parameter $\psi = \{\theta_1, \theta_2, \dots, \theta_G\}$

where $\theta_j = (\pi_j, \mu_j, \Sigma_j) \quad j = 1..G$



density of a mixture of normals (red) and normal components (black)
points from a random sample of the mixture (blue)

Singularities in the likelihood

Maximum likelihood estimation for finite mixture model/clustering model

$x_1, x_1, \ldots, x_n$

Likelihood mixture model

$$\arg\, sup_\theta \sum_{i=1}^{n} log\left(\sum_{j=1}^{k} \pi_j \frac{1}{(2\pi)^{p/2}\left|\Sigma_j\right|^{1/2}} exp\left(-(1/2)(x_i-\mu_j)'\Sigma_j^{-1}(x_i-\mu_j)\right)\right)$$

Likelihood clustering

$$\arg\, sup_\theta \sum_{i=1}^{n}\sum_{j=1}^{k} \log(\pi_j)z_{ij} - (1/2)\log(\left|\Sigma_j\right|)z_{ij} - (1/2)z_{ij}(x_i-\mu_j)'\Sigma_j^{-1}(x_i-\mu_j)$$

Without constraints, the estimation problem is not well posed.

There are singularities in the likelihood

By choosing $\mu_1 = x_1$ and $|\Sigma_1| \longrightarrow 0$, we can get the likelihood goes to ∞.

Then, how we can define Maximum Likelihood Estimator?

Singularities in the likelihood

Maximum likelihood estimation for finite mixture model/clustering model

$x_1, x_1, \dots, x_n$

Likelihood mixture model

$$\arg\, sup_\theta \sum_{i=1}^{n} log\left(\sum_{j=1}^{k} \pi_j \frac{1}{(2\pi)^{p/2}\left|\Sigma_j\right|^{1/2}} exp\left(-(1/2)(x_i - \mu_j)'\Sigma_j^{-1}(x_i - \mu_j)\right)\right)$$

Likelihood clustering

$$\arg\, sup_\theta \sum_{i=1}^{n}\sum_{j=1}^{k} \log(\pi_j) z_{ij} - (1/2)\log(\left|\Sigma_j\right|) z_{ij} - (1/2) z_{ij}(x_i - \mu_j)'\Sigma_j^{-1}(x_i - \mu_j)$$

Without constraints, the estimation problem is not well posed.

There are singularities in the likelihood

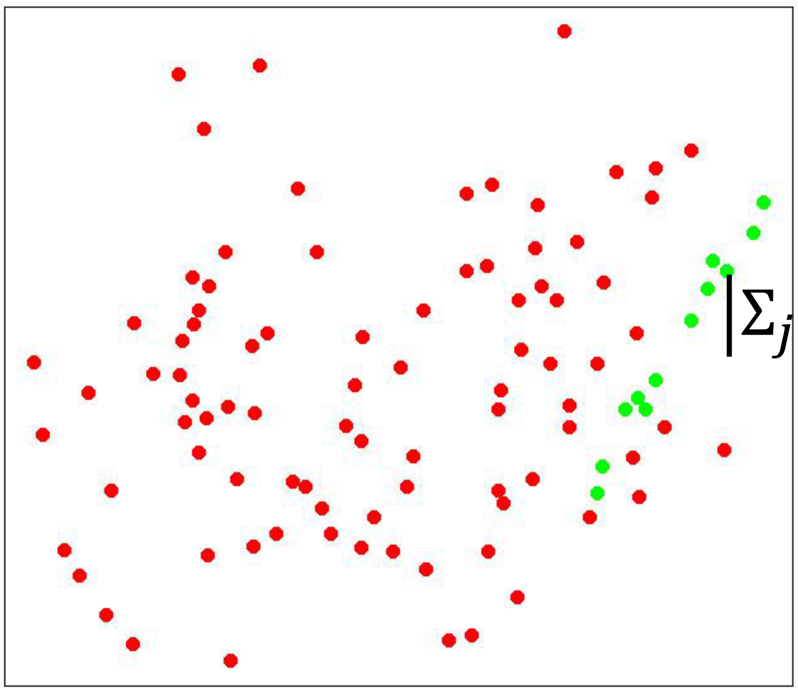
By choosing $\mu_1 = x_1$ and $|\Sigma_1| \longrightarrow 0$, we can get the likelihood goes to ∞.

Then, how we can define Maximum Likelihood Estimator?

maximum of local maximizers??

Singularities in the likelihood

$$\arg sup_{\theta} \sum_{i=1}^{n} log\left(\sum_{j=1}^{k} \pi_j \frac{1}{(2\pi)^{p/2}\left|\Sigma_j\right|^{1/2}} exp\left(-(1/2)(x_i-\mu_j)'\Sigma_j^{-1}(x_i-\mu_j)\right)\right)$$



$\left|\Sigma_j\right|$

Synthetic data set 2 (McP2000)
Mixture of two normal heteroscedastic populations without contamination

Spurious clusters

- "little practical use or real-world interpretation" (McLachlan &Peel, 2000 McP2000)
- "It often seems in these cases that the model is fitting a small localized random pattern in the data rather than any underlying group structure" . (McP2000)

Singularities in the likelihood

Maximum likelihood function for a finite mixture model

How to define MLE?

$$\arg sup_\theta \sum_{i=1}^{n} log\left(\sum_{j=1}^{k} \pi_j \frac{1}{(2\pi)^{p/2}\left|\Sigma_j\right|^{1/2}} exp\left(-(1/2)\left(x_i-\mu_j\right)'\Sigma_j^{-1}\left(x_i-\mu_j\right)\right)\right)$$

To apply constraints

$\Sigma_j = \lambda I$ (k-means // trimmed k-means)

$\Sigma_j = \Sigma$

$\left|\Sigma_j\right| = |\Sigma|$

Hathaway proposal for univariate mixtures (Annals Stat. 1985): In order to get a well posed estimation-problem, to constrain the relative variability between components

$\sigma_i \leq c\sigma_j \quad for\ each\ i,j \;\; 1 \leq i,j \leq k$

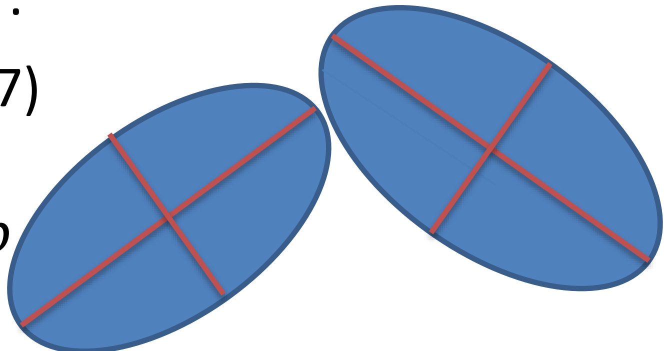
TCLUST. Constraints

In order to get a well posed estimation-problem a solution is to restrict the relative variability between components (Hathaway, Annals Stat. 1985)

For the multivariate case, implemented in TCLUST:

- Eigenvalue constraints (Ingrassia and Rocci, 2007)

$$\frac{\lambda^{l_1}_{\Sigma_{j_1}}}{\lambda^{l_2}_{\Sigma_{j_2}}} \leq c \, , \; 1 \leq j_1, j_2 \leq k \qquad 1 \leq l_1, l_2 \leq p$$



  c is a boundary for the relative variability. It corresponds to restr.fact in R TCLUST.

  To set the boundary for the restrictions equal to c is equivalent to bound the relative size of the tolerance ellipsoids' axis by $\sqrt{c}$

  These constraints are not affine equivariant

- Determinant constraints (McLachlan and Peel, 2000 - McP2000).

  These are affine equivariant constraints

$$\frac{\left|\Sigma_{j_1}\right|}{\left|\Sigma_{j_2}\right|} \leq c \, , \; 1 \leq j_1, j_2 \leq k$$

TCLUST methodology

García-Escudero, Gordaliza, Matrán and M-I (Annals of Stat. 2008).

- Estimator (likelihood based)

$$\arg \sup_{\theta \in R,\, z} \ \sup_{A/\#A = n(1-\alpha)} \sum_{i=1}^{n} \sum_{j=1}^{k} I_A(x_i) z_{ij} \log\left(\pi_j N^p_{\mu_j,\Sigma_j}(x_i)\right)$$

where k is the number of components

$x_1, x_2, \dots, x_n$ is a random sample

$N^p_{\mu,\Sigma}(x)$ is the Normal density

$$\theta = \left(\pi_1, \mu_1, \Sigma_1, \dots, \pi_j, \mu_j, \Sigma_j, \dots, \pi_k, \mu_k, \Sigma_k\right)$$

$$\Sigma_1 = \dots = \Sigma_j = \dots = \Sigma_k = \lambda I_p$$

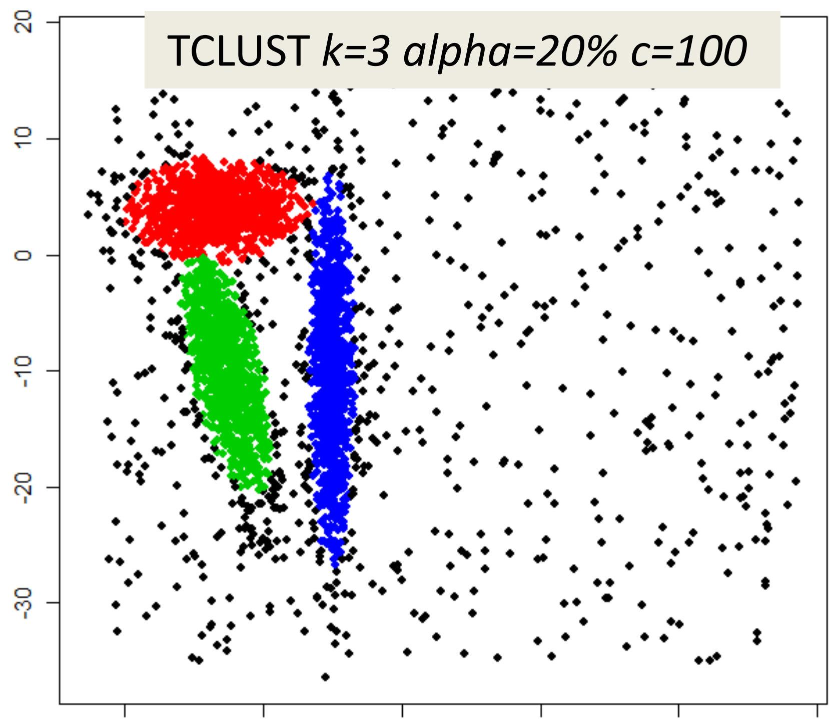
$$\pi_1 = \dots = \pi_j = \dots = \pi_k$$

TCLUST: weaker constraints

- Missing information
  - Membership z_{ij}

    which verifies $\sum_{j=1}^{k} z_{ij} = 1$ & $z_{ij} = 0 \text{ or } 1$
  - Genuineness $I_A(x_i)$

    which verifies $\sum_{i=1}^{n} I_A(x_i) = n(1-\alpha)$

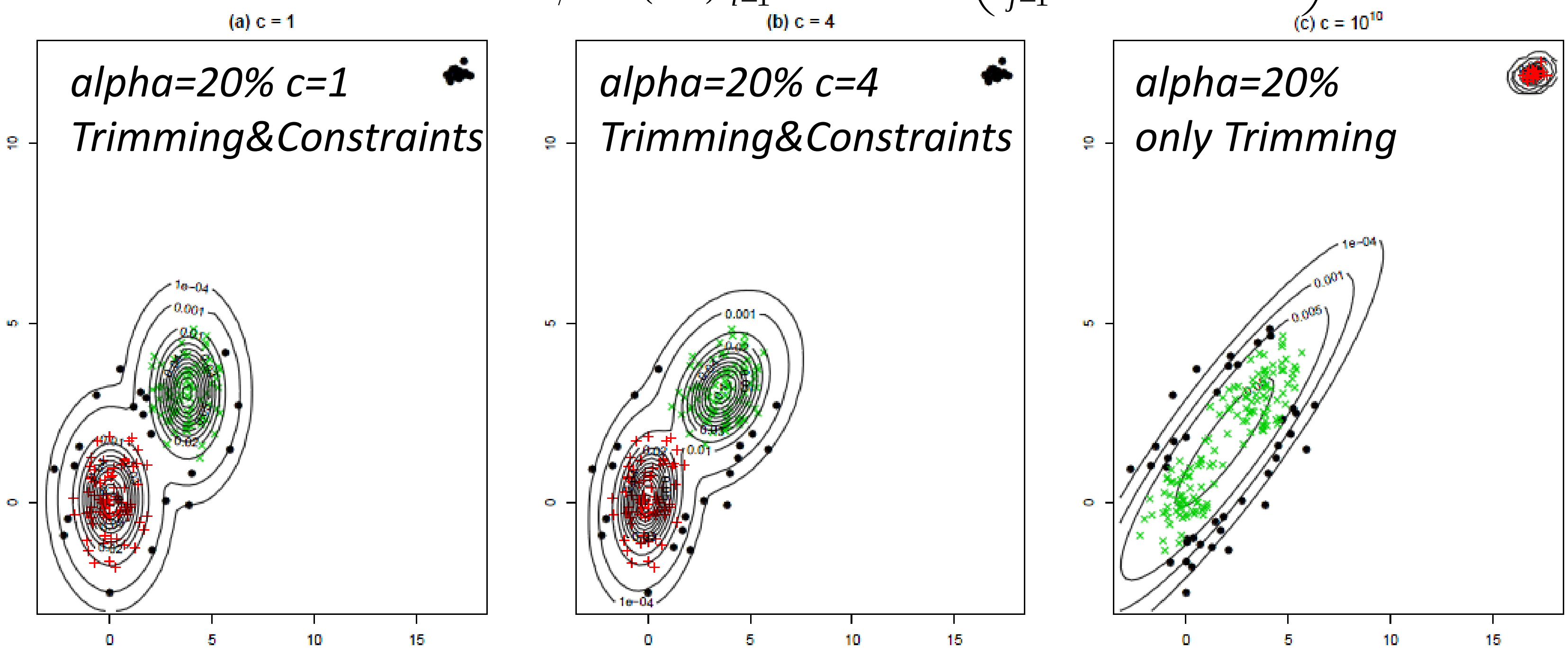


Trimmed & Constrained Maximum likelihood Mixture Normal models

Trimming & Eigenvalue constraints applied to ML finite mixture models estimation

García-Escudero, Gordaliza and M-I (2014). A constrained robust proposal for mixture modeling avoiding spurious solutions. ADAC 8 (1), pp 27-43

$$\arg\sup_{\theta\in R}\ \sup_{A/\#A=n(1-\alpha)} \sum_{i=1}^{n} I_A(x_i)\log\left(\sum_{j=1}^{k} \pi_j N^{p}_{\mu_j,\Sigma_j}(x_i)\right)$$



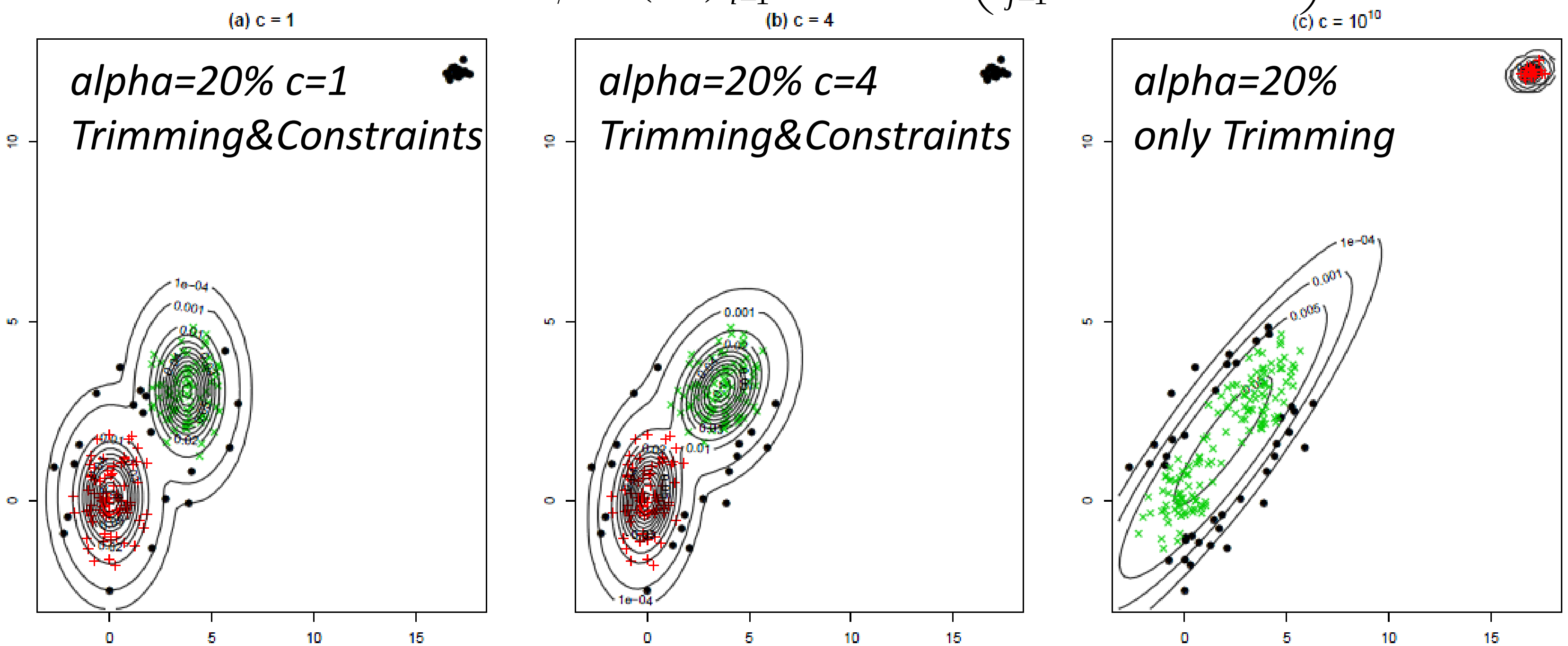
Fitted mixtures for the data in scenario S2 with $n = 200$ when $G = 2$ and $\alpha = 0.2$. Restriction value $c = 1$ is used in (a), $c = 4$ in (b) and $c = 10^{10}$ (almost unrestricted) in (c).

Trimmed & Constrained Maximum likelihood Mixture Normal models

Trimming & Eigenvalue constraints applied to ML finite mixture models estimation

García-Escudero, Gordaliza and M-I (2014). A constrained robust proposal for mixture modeling avoiding spurious solutions. ADAC 8 (1), pp 27-43

$$\arg\sup_{\theta\in R} \sup_{A/\#A=n(1-\alpha)} \sum_{i=1}^{n} I_A(x_i)\log\left(\sum_{j=1}^{k}\pi_j N^{p}_{\mu_j,\Sigma_j}(x_i)\right)$$



Fitted mixtures for the data in scenario S2 with $n = 200$ when $G = 2$ and $\alpha = 0.2$. Restriction value $c = 1$ is used in (a), $c = 4$ in (b) and $c = 10^{10}$ (almost unrestricted) in (c).

Genuine bank notes identification

Bank notes. Flury, B. and Riedwyl, H. (1988). 200 printed Swiss 1000-franc bank notes divided in two groups: 100 genuine and 100 counterfeit notes. It is a well known benchmark data set

Dotto, F., Farcomeni, A., García-Escudero, L. A., & M-I, A. (2018). A reweighting approach to robust clustering. Stat. and Comp., 28(2), 477-493.

Fritz, H., Garcıa-Escudero, L. A., & M-I, A. (2012). tclust: An R package for a trimming approach to cluster analysis. Journal of Statistical Software, 47(12), 1-26.

García-Escudero, L. A., Gordaliza, A., Matrán, C., & M-I, A. (2011). Exploring the number of groups in robust model-based clustering. Stat. & Comp., 21(4), 585-599.

Image from Flury and Riedwyl (1988).

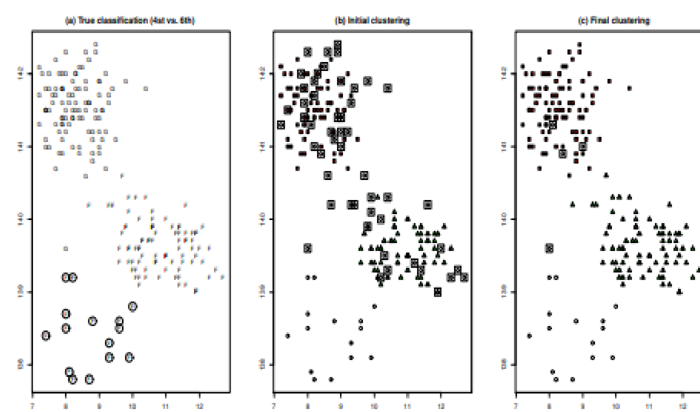
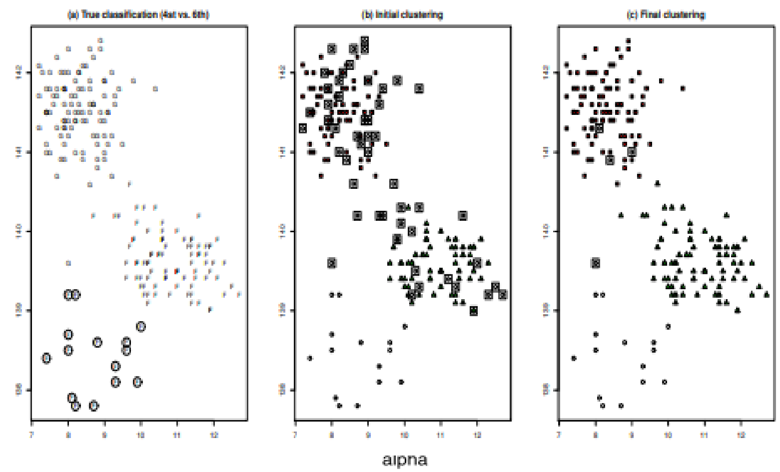


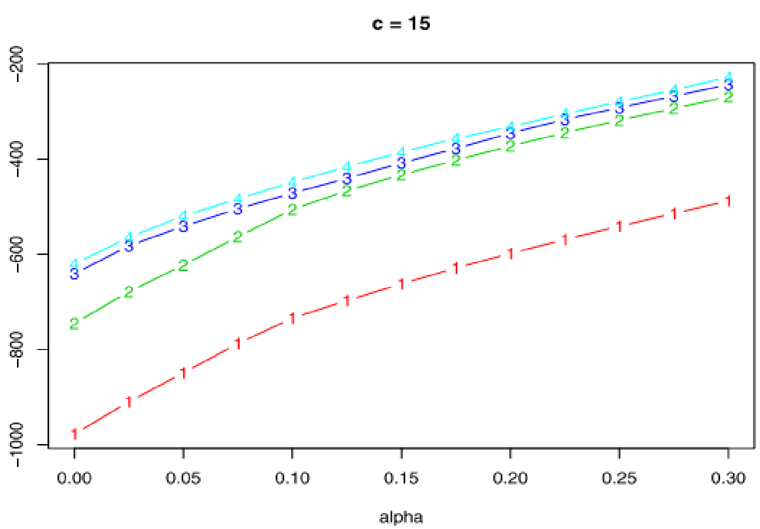
Fig. 8 Fourth against the sixth variable of the Swiss Bank Notes data set. (a) G stands for genuine bills, F for forged ones and 15 bills listed in Flury and Riedwyl (1988) as anomalous ones are surrounded by "○" symbols. (b) The initial TCLUST solution with $\alpha_0 = 0.33$ (c) Final solution when applying the proposed iterative approach. Trimmed observations not coinciding with those in Flury and Riedwyl's list are surrounded by "□" symbols in (b) and (c).

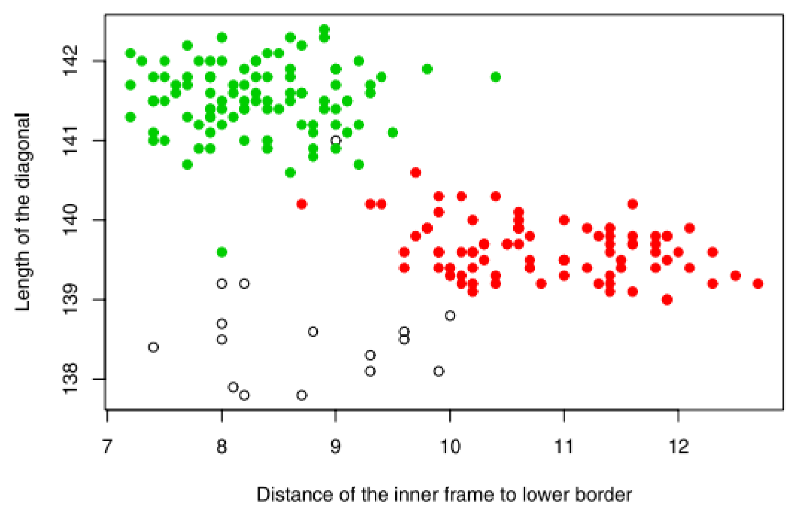
Genuine bank notes identification

Bank notes. Flury, B. and Riedwyl, H. (1988). 200 printed Swiss 1000-franc bank notes divided in two groups: 100 genuine and 100 counterfeit notes. It is a well known benchmark data set

Image from Flury and Riedwyl (1988).







TCLUST. Trimming & Constraints

Early impartial trimming references

- Rousseeuw, P. J., JASA (1984) & Mathematical Statistics and Applications, B , (1985)
- Neykov, N. M. and P. N. Neytchev (1990). Short communications of COMPSTAT
- Gordaliza, A. (1991). Journal of Approximation Theory
- Cuesta-Albertos, J. A., Gordaliza, A., & Matrán, C. (1997). The Annals of Statistics
- Hadi, AS Luceño (1997). Computational Statistics & Data Analysis
- Vandev, D. L., & Neykov, N. M. (1998). A Journal of Theoretical and Applied Statistics
- García-Escudero, Gordaliza, Matrán and M-I. (2008) Annals of Statistics

Early references related with constraints application proposals

- Hathaway (1985). Annals of Statistics
- Gallegos and Ritter (2005). Annals of Statistics
- Ingrassia and Rocci (2007). Computational Statistics & Data Analysis
- García-Escudero, Gordaliza, Matrán and M-I. (2008) Annals of Statistics

TCLUST
Statistical properties

Statistical methodology

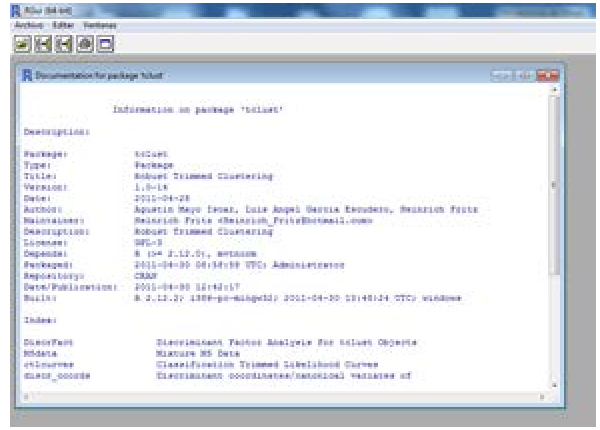
- Well posed statistical problem. We are interested in the maximum in the restricted parameter space.
- Existence and Consistency (García-Escudero, Gordaliza, Matrán and M-I, Annals of Stat. 2008).
- Breakdown point≈alpha (in the sense of Hennig (2004))

TCLUST
Algorithm

Fast algorithm for TCLUST (Fritz, García-Escudero and M-I. CSDA,2013)

- Random starts
- Iterations
  - E step, to assign each point to the closest component, in the sense given by a greatest value in the discriminant functions $\pi_j N_{\mu_j,\Sigma_j}(x_i)$ and to obtain the **optimal *A* set** is given by the 1-α proportion of closest points to the model in the sense of $\pi_{j^{opt}} N_{\mu_{j^{opt}},\Sigma_{j^{opt}}}(x_i)$
  - M step, to obtain the best value for the parameters in **the constrained space**. The current release of the algorithm reduces this search to a set of $kp+2$ possible solutions obtained in a explicit way. In relation with the classical ML estimator, the change appears in the estimation of Σ_j , which corresponds to the projection of matrices S_j in the constrained space.

TCLUST
R & MATLAB



TCLUST in CRAN. TCLUST package. Maintainer: Valentin Todorov

Fritz, García-Escudero and M-I (2012) Tclust: An R Package for a Trimming Approach to Cluster Analysis. J.Stat. Soft. *47*(12), 1-26.

TCLUST in Matlab. FSDA library.

Riani, Perrotta & Torti, F. 2012. FSDA: A MATLAB toolbox for robust analysis and interactive data exploration. *Chemometrics and Intelligent Laboratory Systems*, 116, 17–32

Ro.Sta.Bi.Da.C - CENTRO DI STATISTICA ROBUSTA PER GRANDI BANCHE DATI (ROBUST STATISTICS FOR BIG DATA CENTRE) of University of Parma. Marco Riani and Andrea Cerioli.

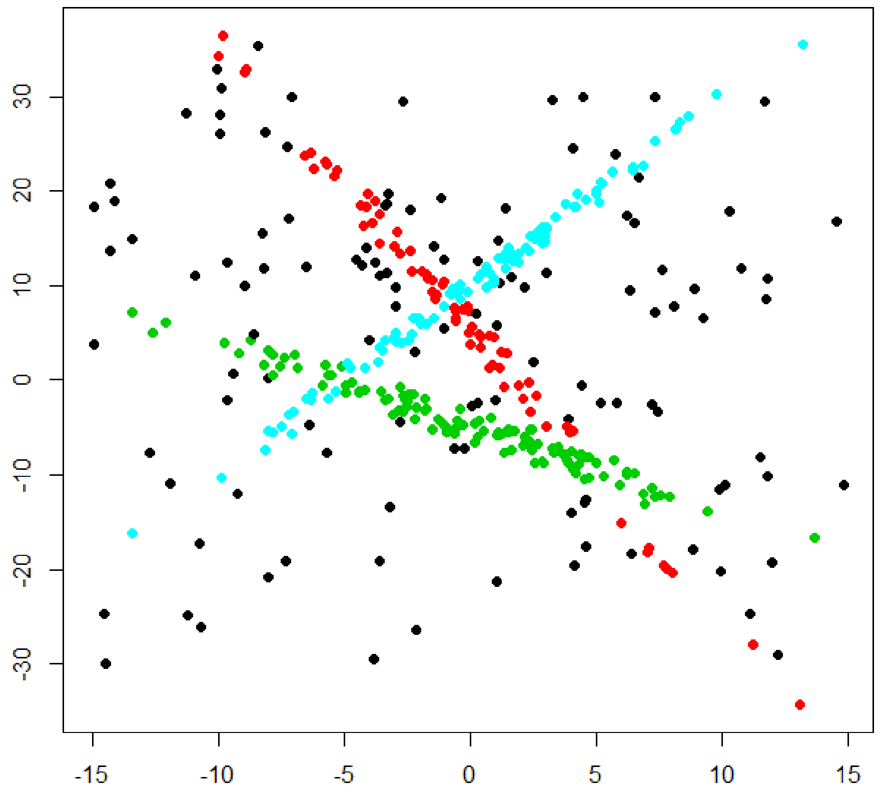
JRC Ispra. Domenico Perrotta and Francesca Torti.

Clustering/mixture regression models

Trimming & Constraints
Clustering of regression models

Trimming & Constraints in order to get robust clustering of regression models. García-Escudero, Gordaliza, M-I & San Martín (CSDA, 2010).

$$\arg \sup_{\theta \in R} \sup_{z} \sup_{A/\#A=n(1-\alpha)} \sum_{i=1}^{n} \sum_{j=1}^{k} I_A(x_i, y_i) z_{ij} \log\left(\pi_j N^1_{0,\sigma_j}\left(y_i - \beta_0 - \beta_j x_i\right)\right)$$



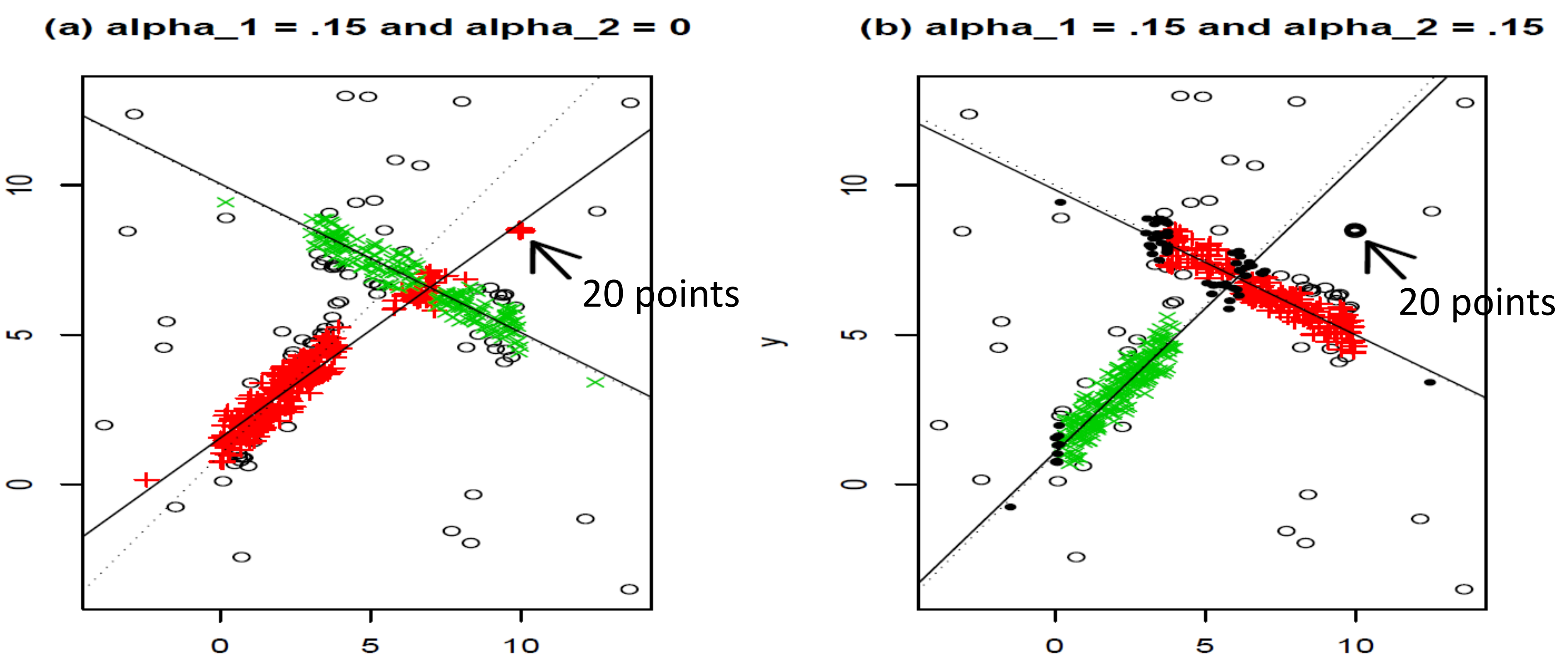
$$\frac{\sigma^2_{j_1}}{\sigma^2_{j_2}} \le c\ ,\ 1 \le j_1, j_2 \le k$$

alpha=30% c=10

Trimming & Constraints
Clustering of regression models

Trimming & Constraints in order to get robust clustering of regression models. García-Escudero, Gordaliza, M-I & San Martín (CSDA, 2010).

A second trimming can be included in the E step of EM algorithm in order to eliminate outliers in explanatory variables.

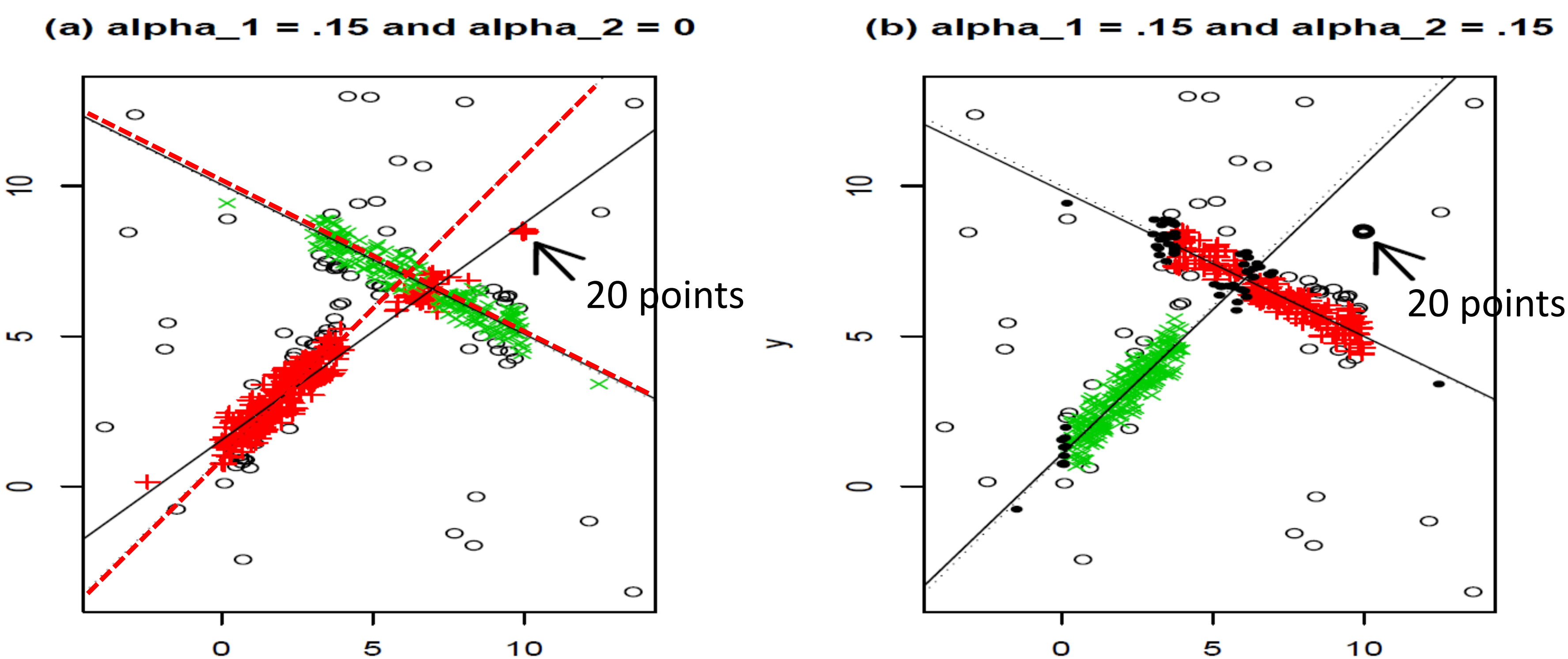


Example with artificial data in García-Escudero Gordaliza, M-I & San Martín (CSDA, 2010).

Trimming & Constraints
Clustering of regression models

Trimming & Constraints in order to get robust clustering of regression models. García-Escudero, Gordaliza, M-I & San Martín (CSDA, 2010).

A second trimming can be included in the E step of EM algorithm in order to eliminate outliers in explanatory variables.



Example with artificial data in García-Escudero Gordaliza, M-I & San Martín (CSDA, 2010).

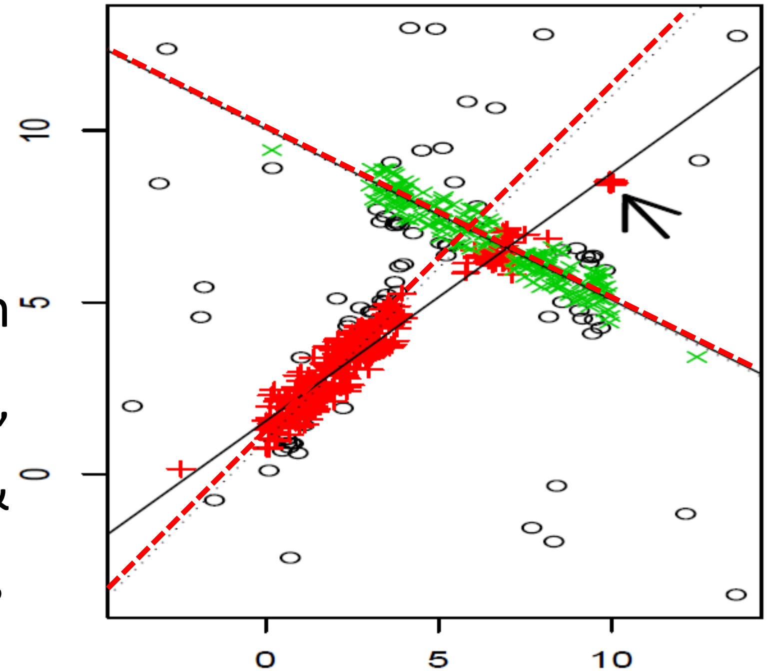
Cluster weighted model

Cluster Weighted Model (CWM) is a mixture approach to modeling the joint probability of data coming from a heterogeneous population. Introduced in Gershenfeld (1997) under Gaussian and linear assumptions.

CWM decomposes the joint probability in each component of the mixture as the product of the marginal and the conditional distributions.

$$\sum_{j=1}^{k} \pi_j N^1_{0,\sigma_j}\left(y - \beta_0 - \beta_j x\right) N^p_{\mu,\Sigma_j}(x)$$

Ingrassia et al. (2012) shows that Gaussian CWM includes, as special cases, Multivariate Finite Mixture Models & Classical Finite Mixture Regression Models



Trimming & Constraints
Cluster weighted model

Trimmed Cluster Weighted Restricted Modeling (TCWRM). García-Escudero, Gordaliza, Greselin, Ingrassia & M-I (Stat&Comp, 2017)

$$\arg\sup_{\theta\in R}\ \sup_{A/\#A=n(1-\alpha)} \sum_{i=1}^{n} I_A(x_i, y_i)\log\left(\sum_{j=1}^{k}\pi_j N^1_{0,\sigma_j}\left(y_i-\beta_0-\beta_j x_i\right)N^p_{\mu,\Sigma_j}(x_i)\right)$$

We apply jointly trimming and two kind of constraints

- Eigenvalue constraints for controlling the relative variability of regression errors

$$\frac{\sigma^2_{j_1}}{\sigma^2_{j_2}} \le c_x\ ,\ 1\le j_1, j_2 \le k$$

- Eigenvalue constraints for controlling the relative variability of explanatory variables

$$\frac{\lambda^{l_1}_{\Sigma_{j_1}}}{\lambda^{l_2}_{\Sigma_{j_2}}} \le c_\varepsilon\ ,\ 1\le j_1, j_2 \le k \qquad 1\le l_1, l_2 \le p$$

Trimming & Constraints
Cluster weighted model

Trimmed Cluster Weighted Restricted Modeling (TCWRM). García-Escudero, Gordaliza, Greselin, Ingrassia & M-I (Stat&Comp, 2017)

$$\arg\sup_{\theta \in R} \sup_{A/\#A=n(1-\alpha)} \sum_{i=1}^{n} I_A(x_i, y_i) \log\left(\sum_{j=1}^{k} \pi_j N^1_{0,\sigma_j}\left(y_i - \beta_0 - \beta_j x_i\right) N^p_{\mu,\Sigma_j}(x_i) \right)$$

Trimming & Restrictions in CWM fitting (TCWRM)

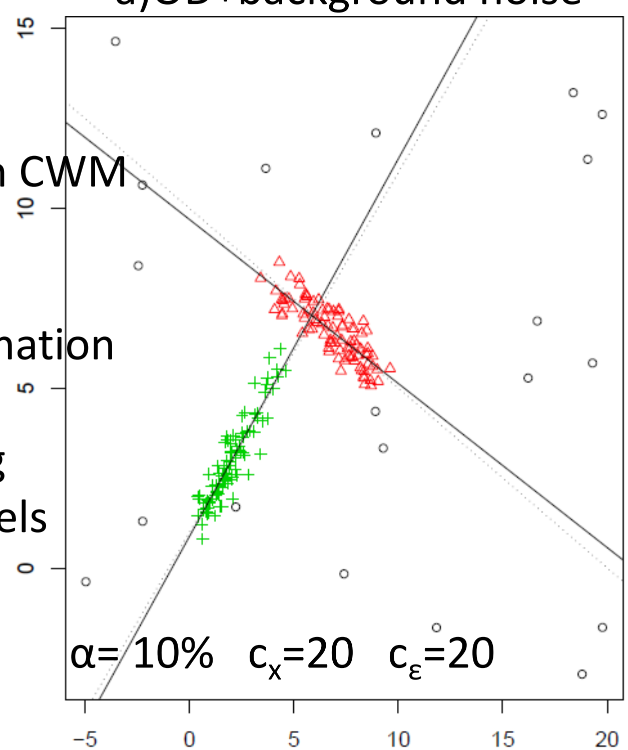
a) OD+background noise

b) OD+pointwise contamination

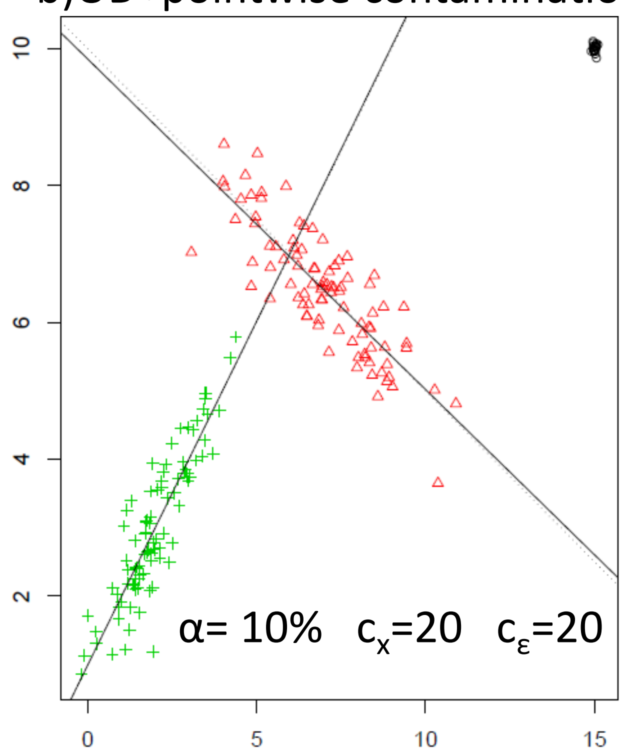
Black lines: TCWRM fitting

Dot lines: true linear models

a)OD+background noise



b)OD+pointwise contamination



From García-Escudero, Gordaliza, Greselin, Ingrassia & M-I (2014)

TCLUST REG // TCLUST CWM

Torti, F., Perrotta, D., Riani, M., & Cerioli, A. (2019). Assessing trimming methodologies for clustering linear regression data. *Advances in Data Analysis and Classification*, *13*(1), 227-257.

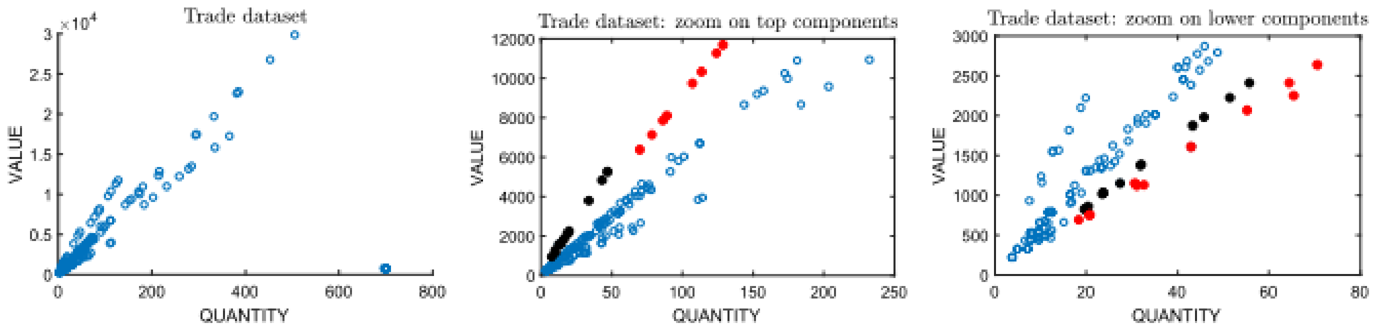


Fig. 12 Scatterplots of case study 5. Trade dataset formed by customs declarations made by an EU importer. The axes report the declared values (y-axis) and quantities (x-axis). The left panel plots the data in the original scale. The central and right panels zoom in the data to highlight the presence of components that are difficult to notice in the original scale

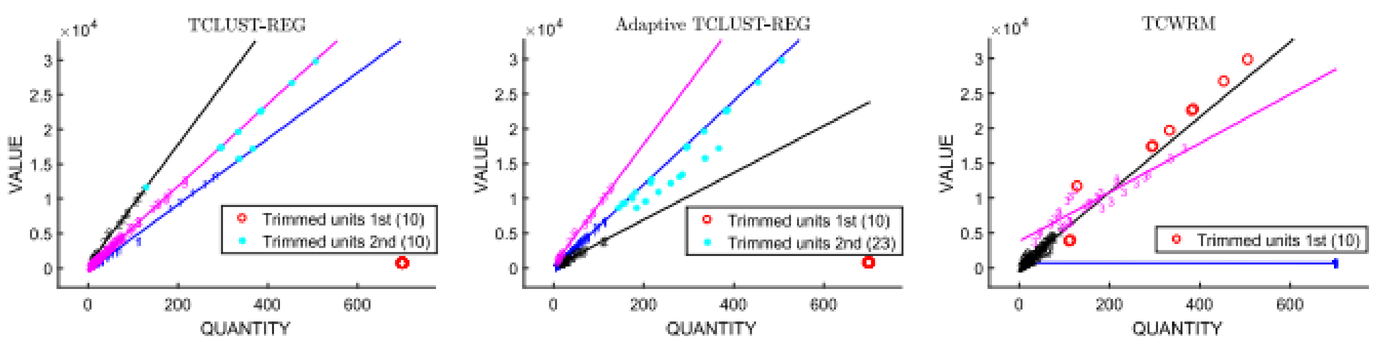


Fig. 13 Case study 5. Dataset of Figure 12 analyzed with three components ($G = 3$) with, from left to right, TCLUST-REG, Adaptive TCLUST-REG and TCWRM

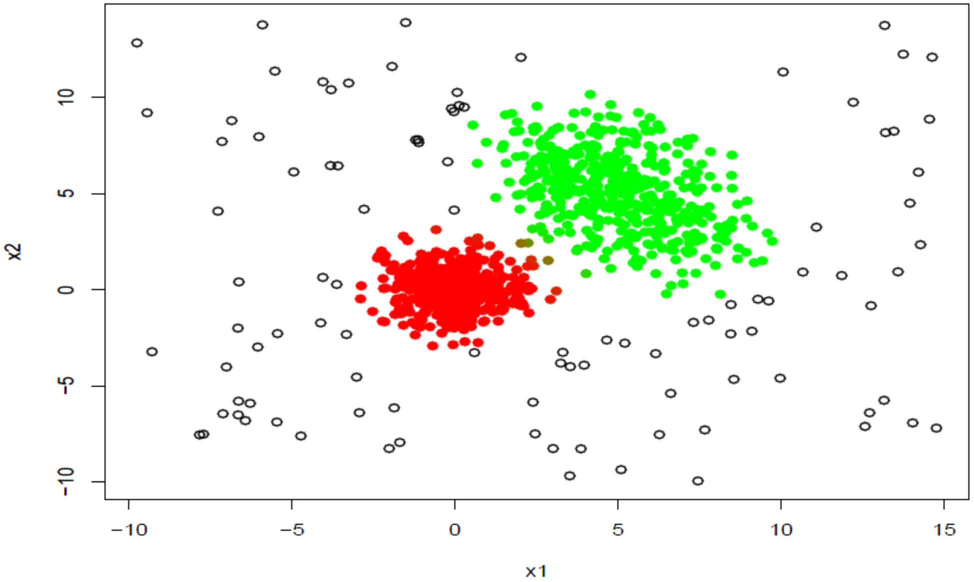
Fuzzy TCLUST

Trimming & Constraints Fuzzy clustering

Fritz, García-Escudero and M-I (2013), Robust Constrained Fuzzy Clustering. Information Sciences, 245, 38-52

$$\arg\sup_{\mu,u} \sup_{A/P_n(A)=1-\alpha} \sum_{i=1}^{n} I_A(x_i) \sum_{j=1}^{k} u_{ij}^m \log\left(\pi_j N_{(\mu_j,\Sigma_j)}(x_i)\right)$$

$$\frac{\lambda_{\Sigma_{j_1}}^{l_1}}{\lambda_{\Sigma_{j_2}}^{l_2}} \le c\,,\ 1 \le j_1, j_2 \le k \qquad 1 \le l_1, l_2 \le p$$



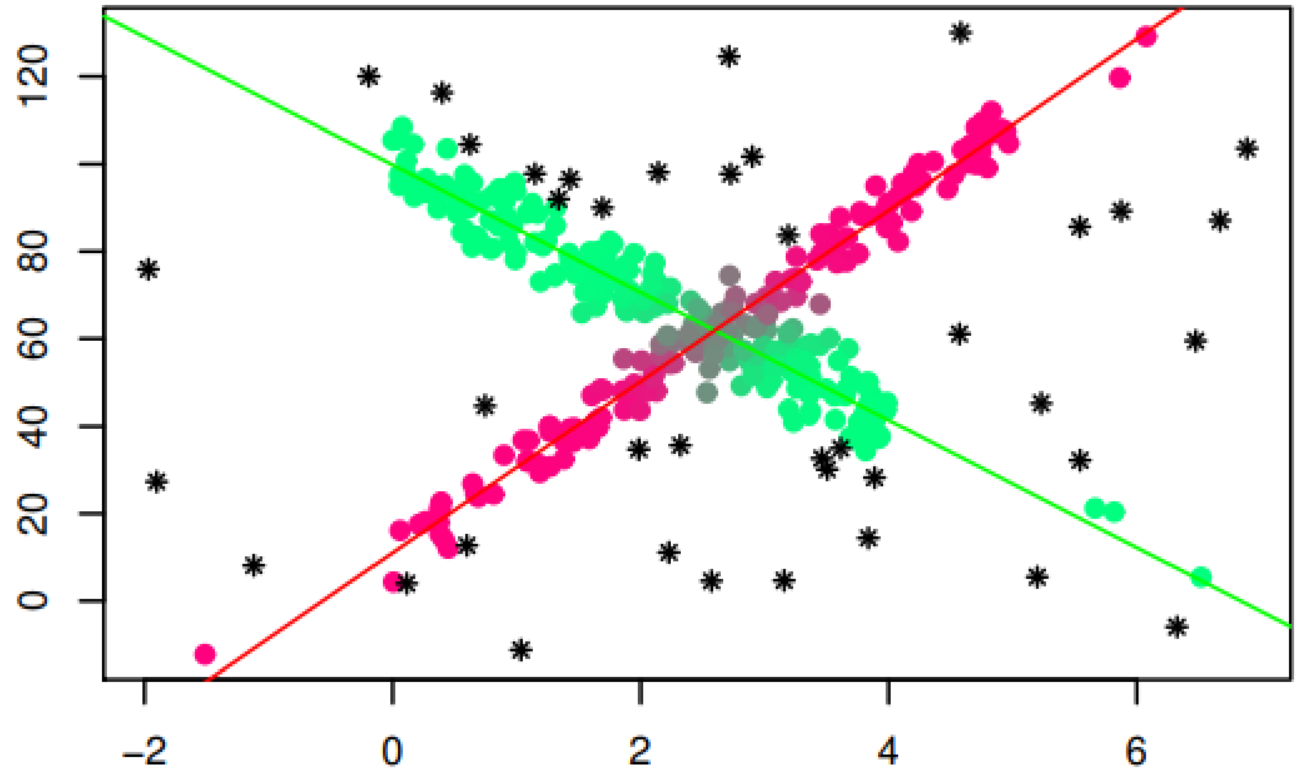
$\alpha = 0.1$ for a 10% contamination level $\Rightarrow$ Trimmed points: "o"

Fuzzy TCLUST Reg

Dotto, Farcomeni, García-Escudero, M-I (2017) A fuzzy approach to robust regression clustering. ADAC

$$\arg\sup_{\theta \in R} \sup_{z} \sup_{A/\#A=n(1-\alpha)} \sum_{i=1}^{n} \sum_{j=1}^{k} I_A(x_i, y_i) u_{ij}^{m} \log\left(\pi_j N^1_{0,\sigma_j}\left(y_i - \beta_0 - \beta_j x_i\right)\right)$$

$$\frac{\sigma^2_{j_1}}{\sigma^2_{j_2}} \leq c\,,\; 1 \leq j_1, j_2 \leq k$$



Robust fuzzy cluster weighted modeling

Trimming & Constraints to Fuzzy cluster weighted Model

$$\arg\sup_{\theta \in R, u} \sup_{A / \#A = n(1-\alpha)} \sum_{i=1}^{n} I_A(x_i, y_i) \left(\sum_{j=1}^{k} u_{ij}^m \log\left(\pi_j N_{0,\sigma_j}^1 (y_i - \beta_0 - \beta_j x_i) N_{\mu,\Sigma_j}^p (x_i)\right)\right)$$

We apply jointly trimming and two kind of constraints

- Eigenvalue constraints for controlling the relative variability of regression errors

$$\frac{\sigma_{j_1}^2}{\sigma_{j_2}^2} \le c_x \ , \ 1 \le j_1, j_2 \le k$$

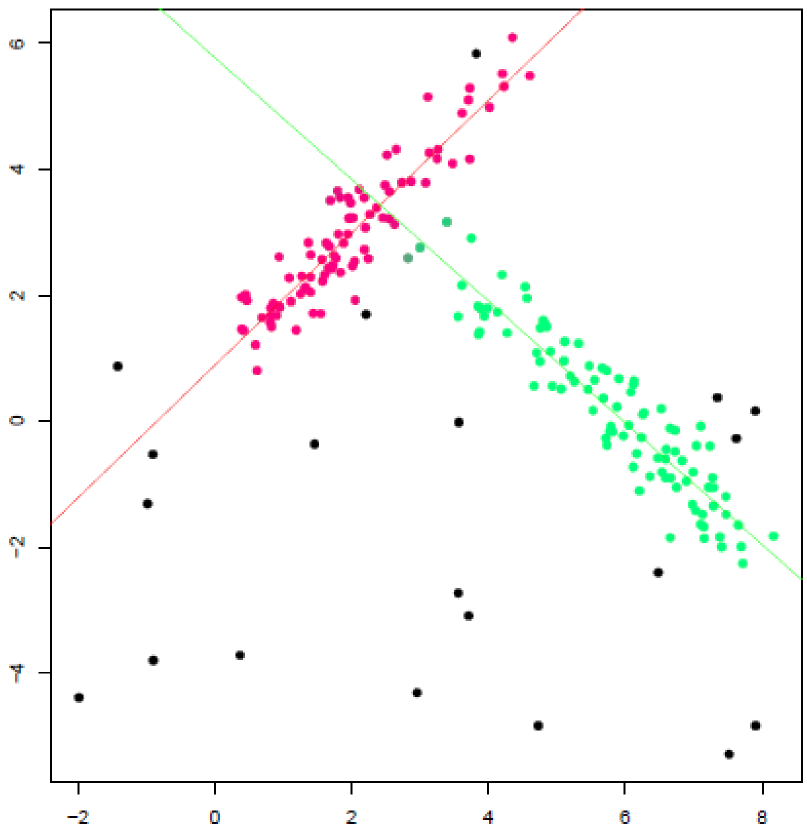
- Eigenvalue constraints for controlling the relative variability of explanatory variables

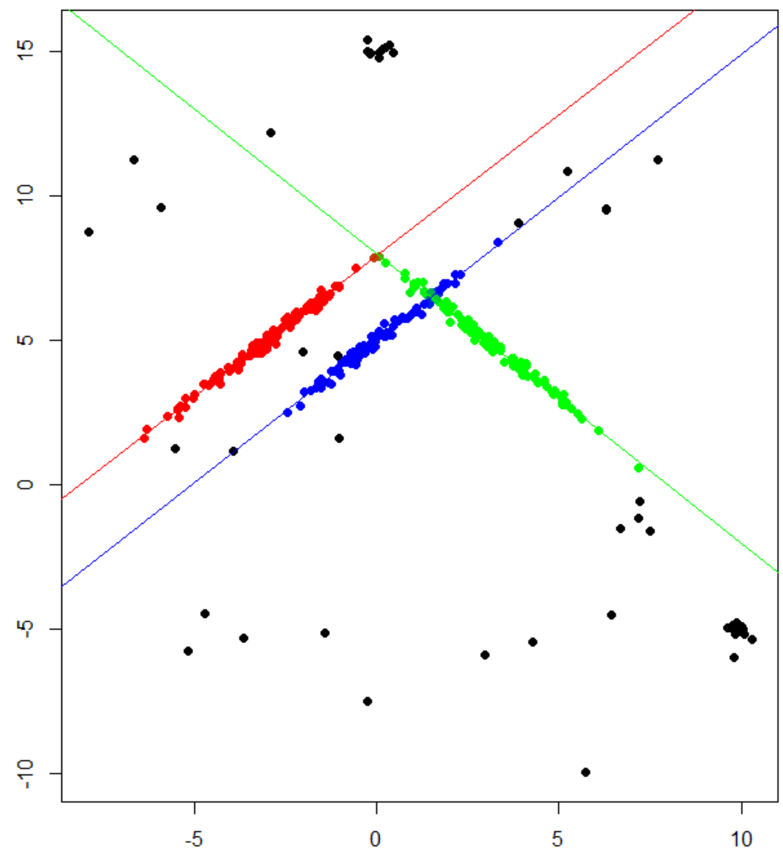
$$\frac{\lambda_{\Sigma_{j_1}}^{l_1}}{\lambda_{\Sigma_{j_2}}^{l_2}} \le c_\varepsilon \ , \ 1 \le j_1, j_2 \le k \qquad 1 \le l_1, l_2 \le p$$

Robust fuzzy cluster weighted modeling

Trimming & Constraints to Fuzzy cluster weighted Model

$$\arg\sup_{\theta\in R,u}\ \sup_{A/\#A=n(1-\alpha)} \sum_{i=1}^{n} I_A(x_i,y_i)\left(\sum_{j=1}^{k} u_{ij}^{m}\log\left(\pi_j N^{1}_{0,\sigma_j}\left(y_i-\beta_0-\beta_j x_i\right) N^{p}_{\mu,\Sigma_j}(x_i)\right)\right)$$



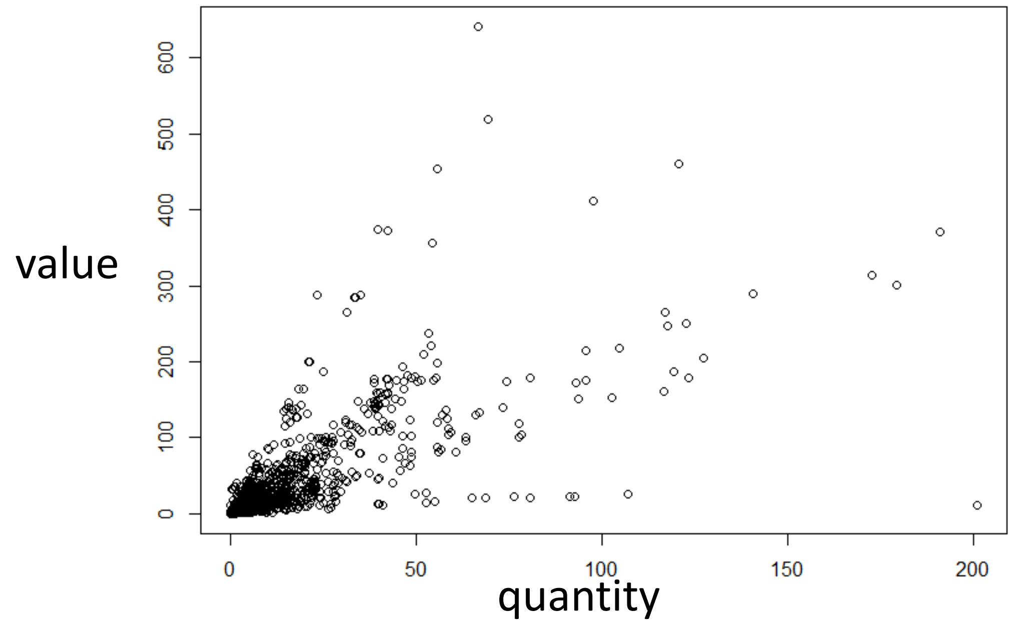


ComExt database

The ComExt Extra-European trade database provides statistics on merchandise trade among European Union member states, and between member states and global partners. ComExt, published by Eurostat, is based on data provided by the statistical agencies of the EU member states and trading partners. The statistics of interest for anti-fraud are mainly the traded **volumes and values for a fixed product**, which are aggregated monthly by Eurostat.

We are interested in applying robust clustering procedures for identifying outliers in the ComExt Extra-European trade database by thinking about its usefulness in fraud detection.

There are robust procedures available for clustering data in different settings, including ones devoted to identifying clusters around linear subspaces which appear to be well suited for datasets in this database.



ComExt database

Cerioli, A. & Perrotta, D. (2014) Robust clustering around regression lines with high density regions. Adv Data Anal Classif 8, 5-26

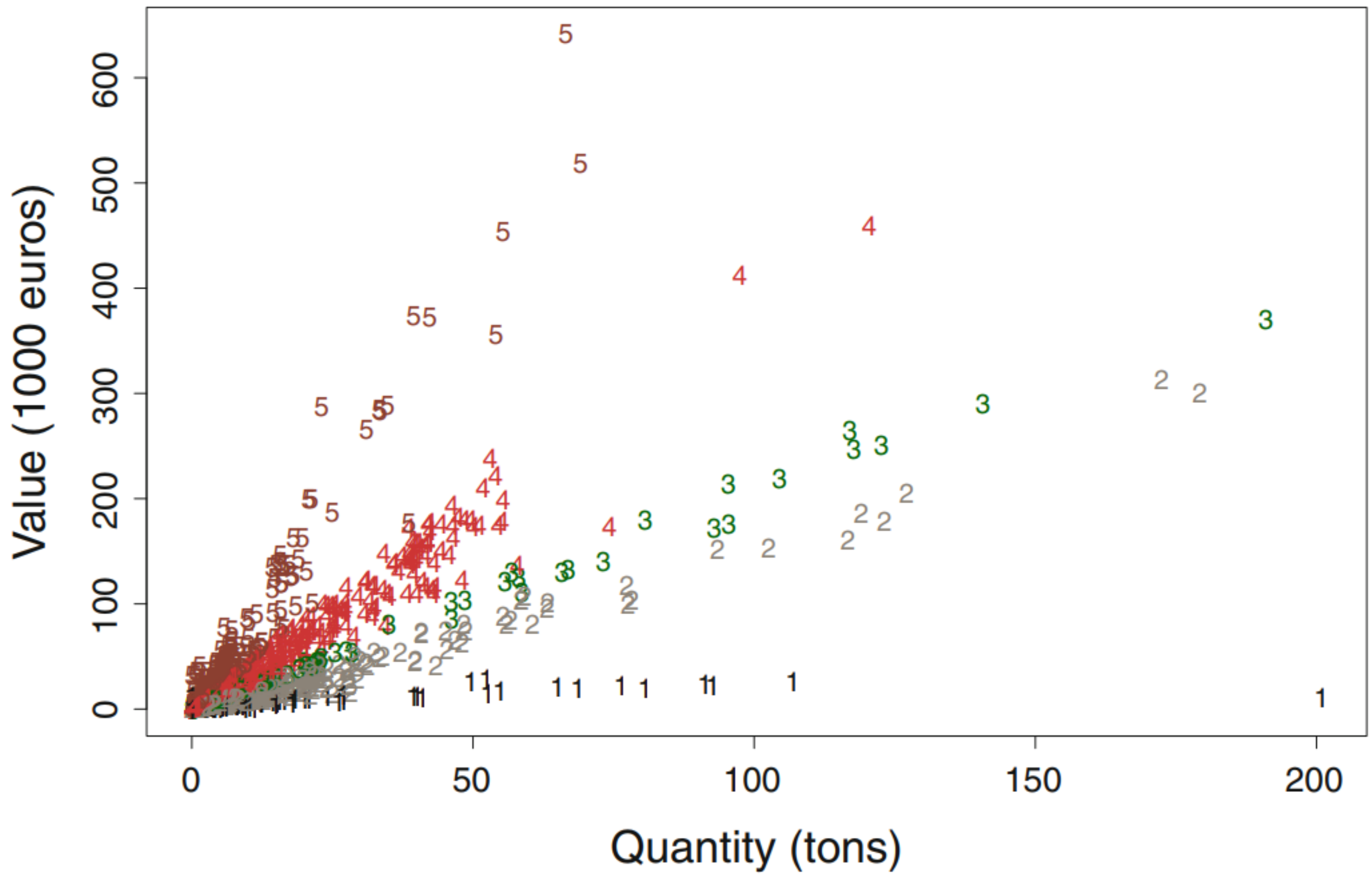
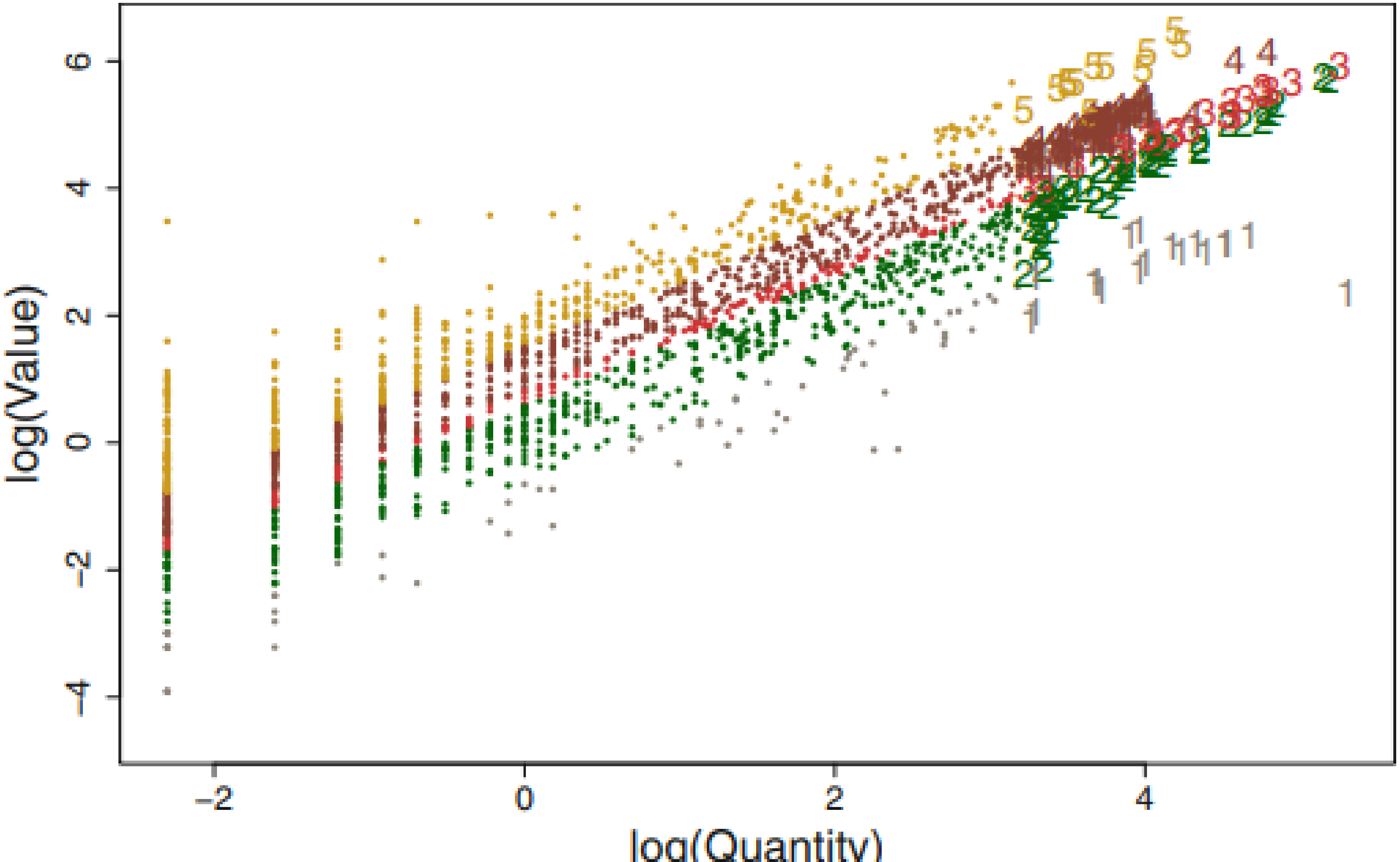


Fig. 1 Spices data set: scatter plot of value (y) and quantity (x), together with cluster membership assigned

ComExt database

Cerioli, A. & Perrotta, D. (2014) Robust clustering around regression lines with high density regions. Adv Data Anal Classif 8, 5-26

log-log plot



ComExt database

Cerioli, A. & Perrotta, D. (2014) Robust clustering around regression lines with high density regions. Adv Data Anal Classif 8, 5-26

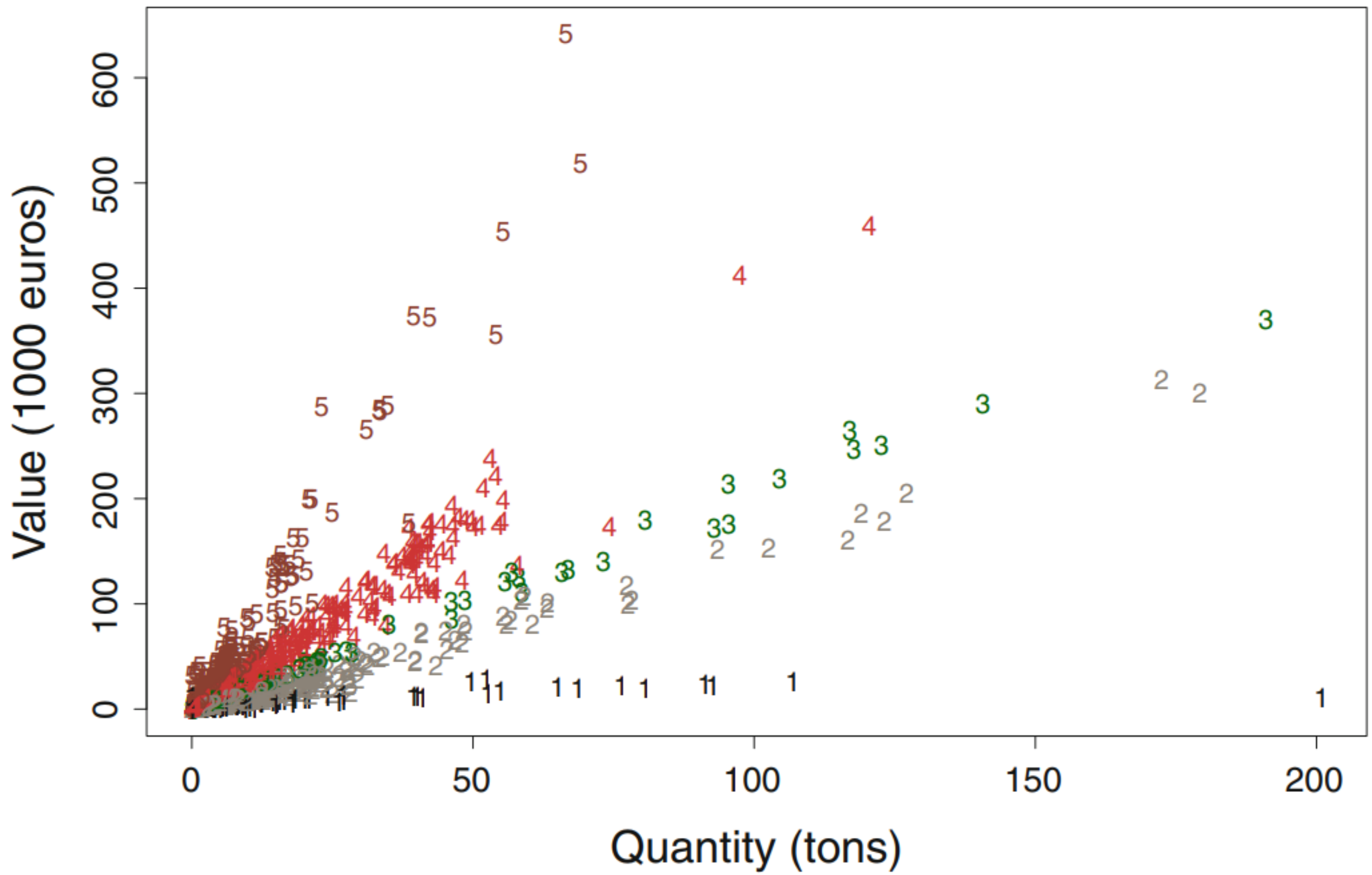
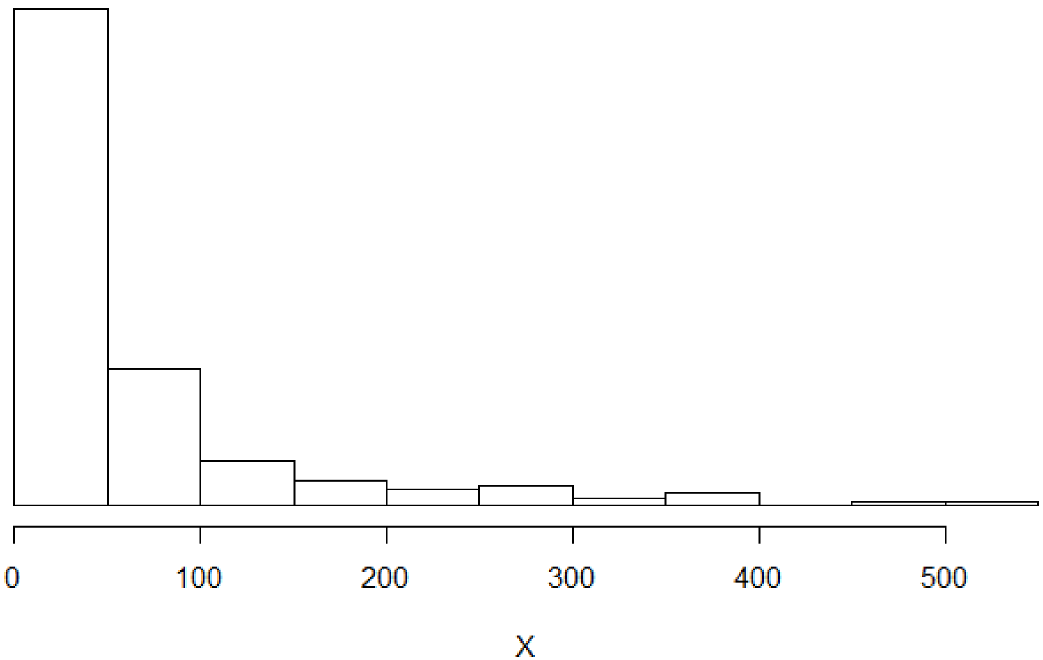


Fig. 1 Spices data set: scatter plot of value (y) and quantity (x), together with cluster membership assigned

tabase

stering around regression lines with

8, 5-26



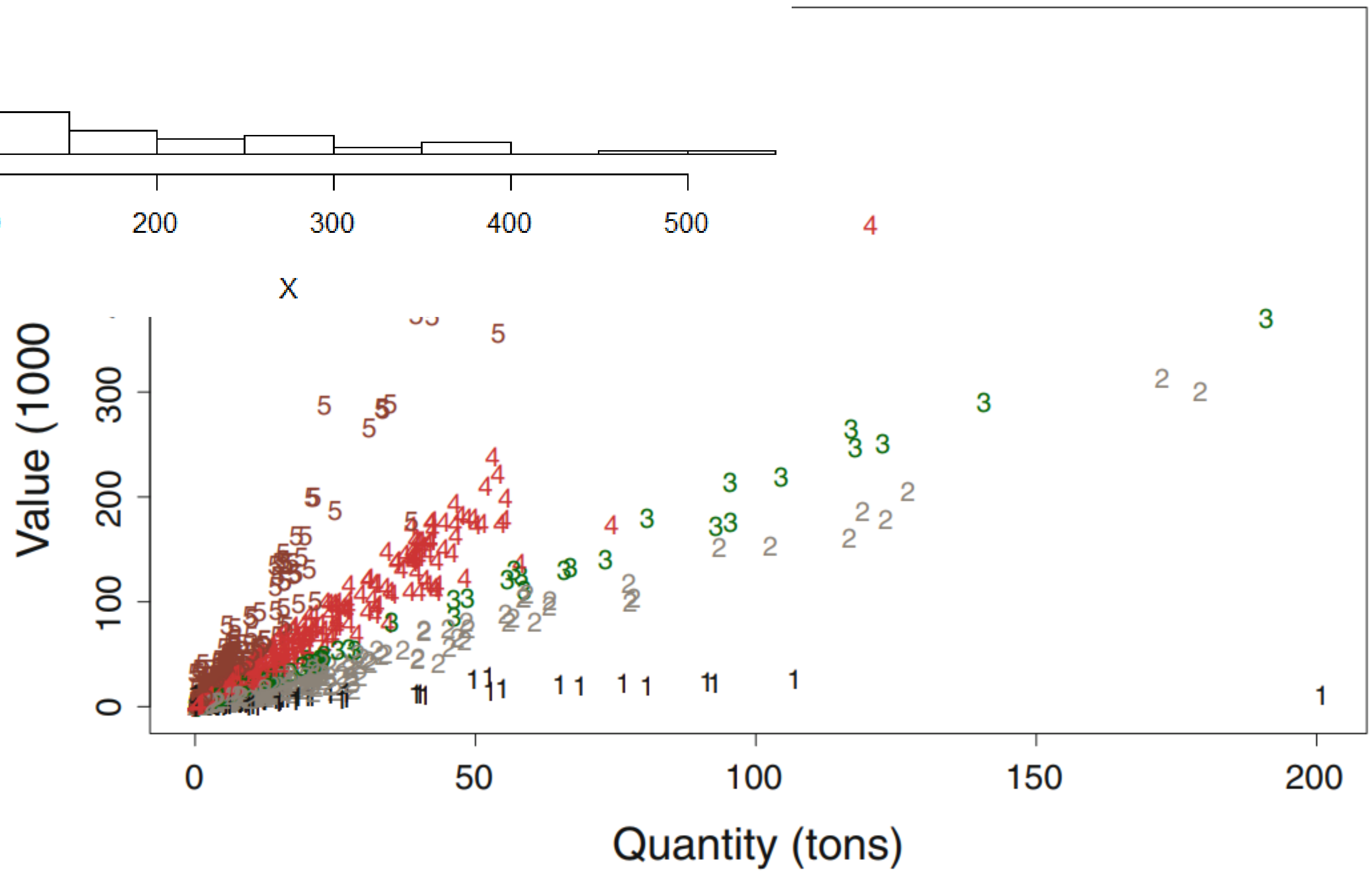
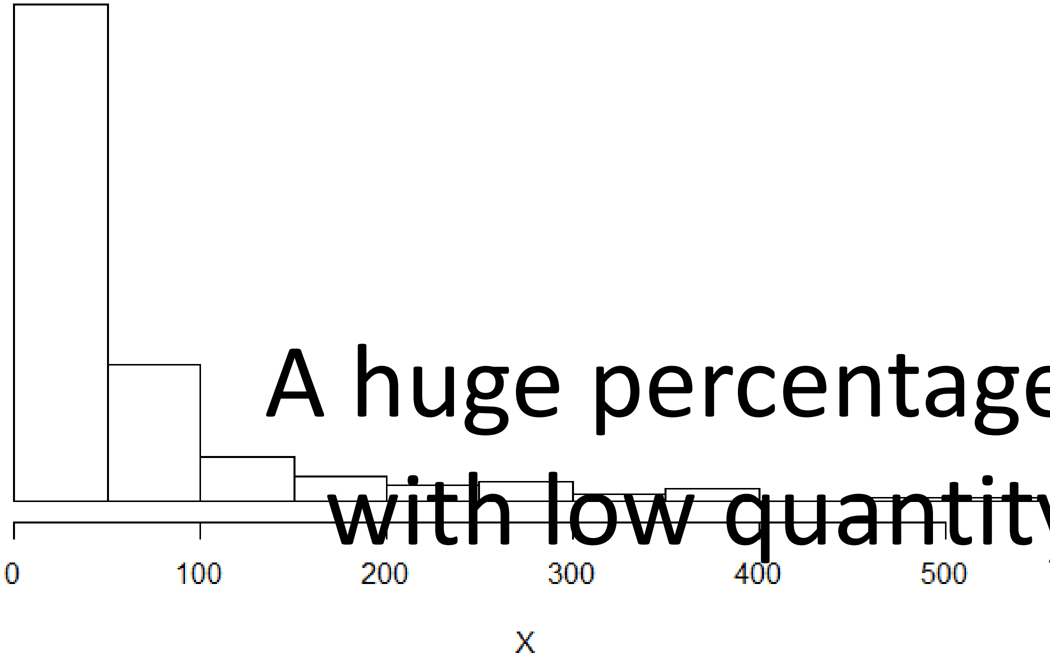


Fig. 1 Spices data set: scatter plot of value (y) and quantity (x), together with cluster membership assigned

tabase

stering around regression lines with

8, 5-26



A huge percentage of observations with low quantity and low value

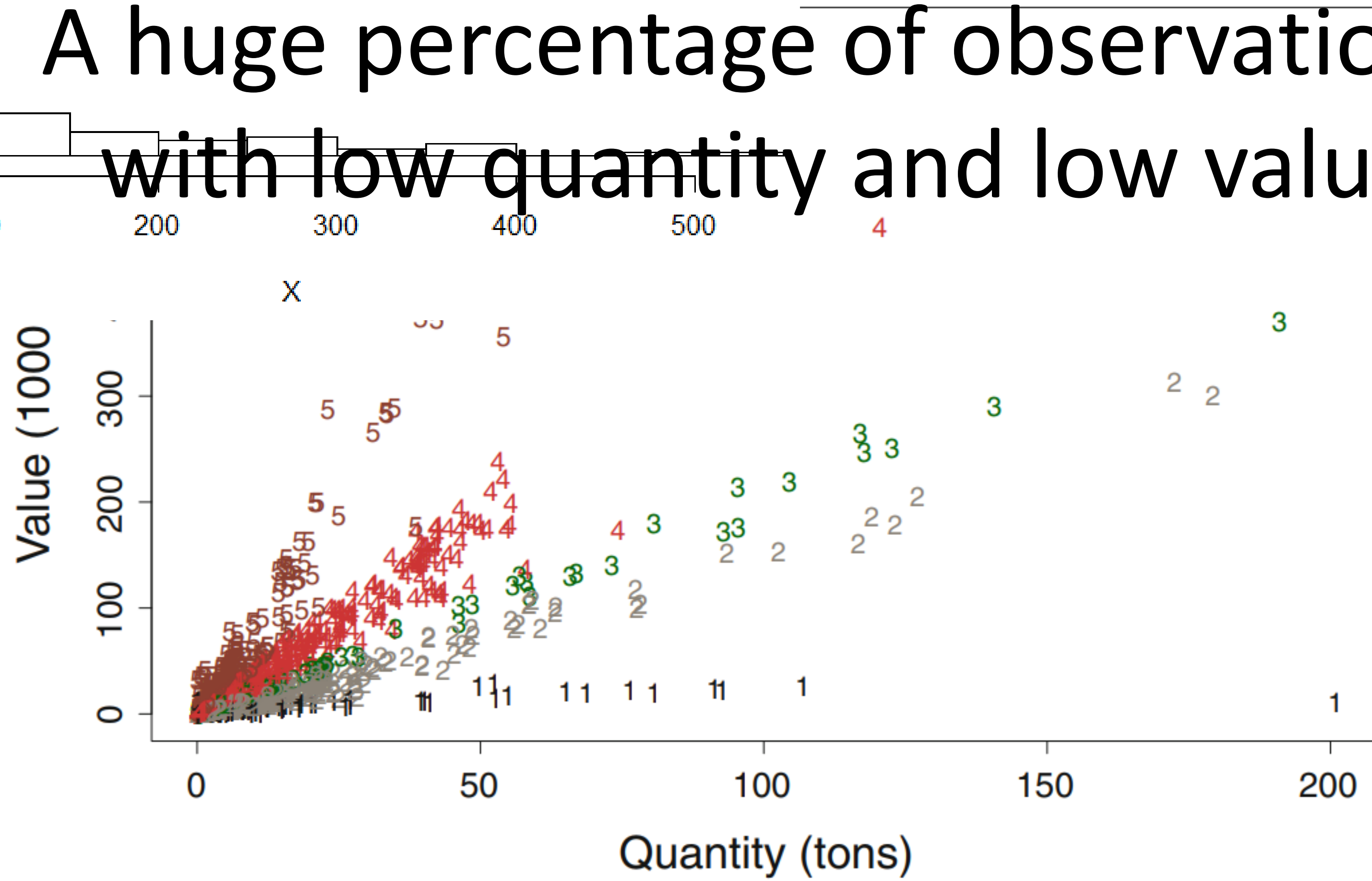
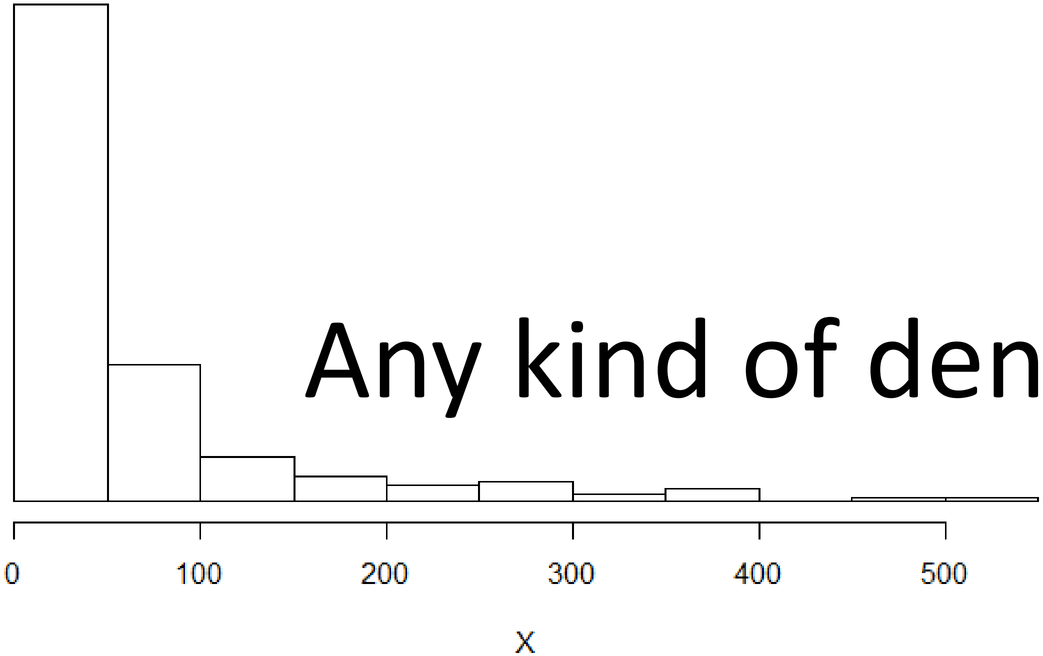


Fig. 1 Spices data set: scatter plot of value (y) and quantity (x), together with cluster membership assigned

tabase

stering around regression lines with

8, 5-26



Any kind of denoising is needed

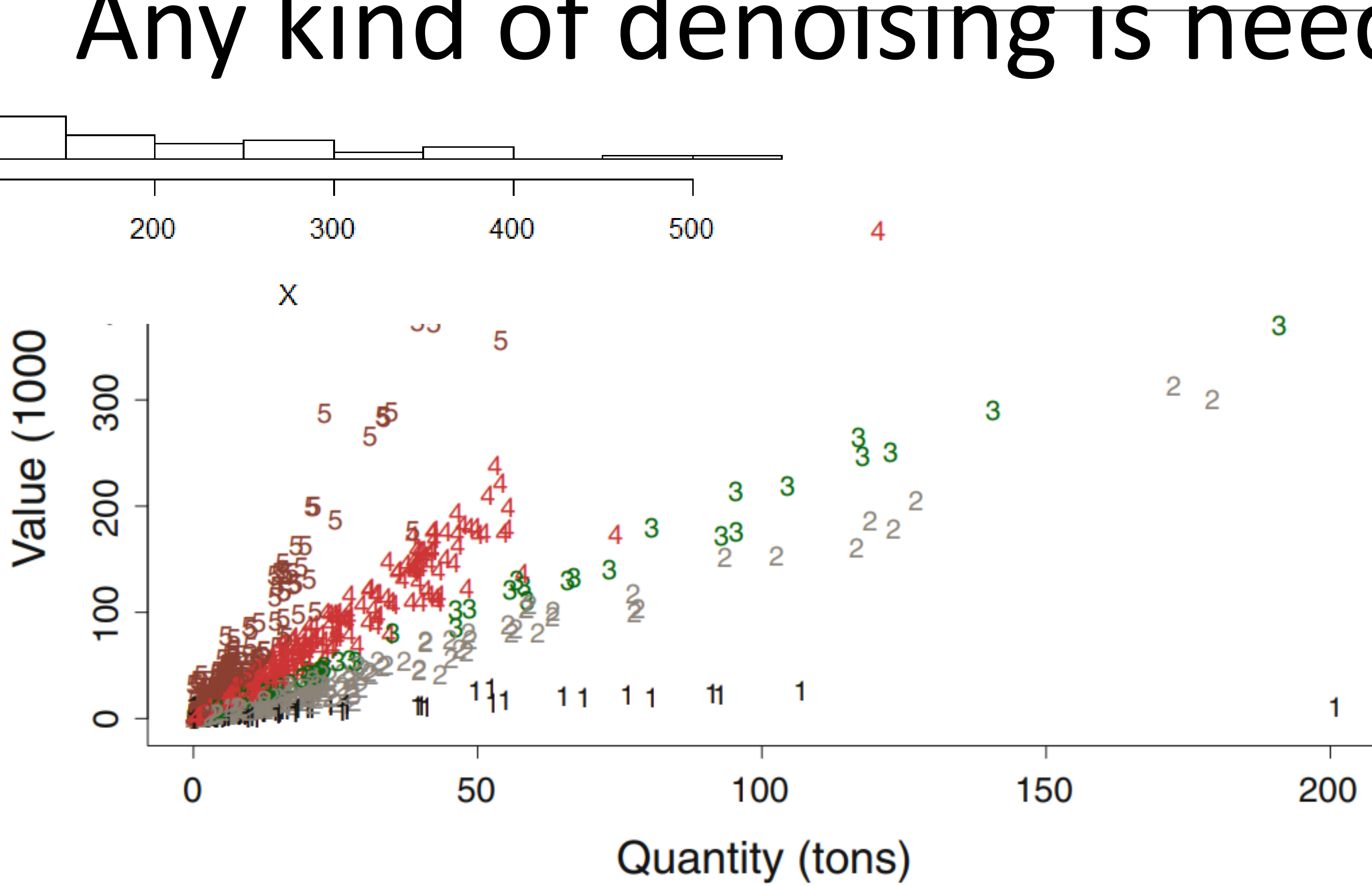
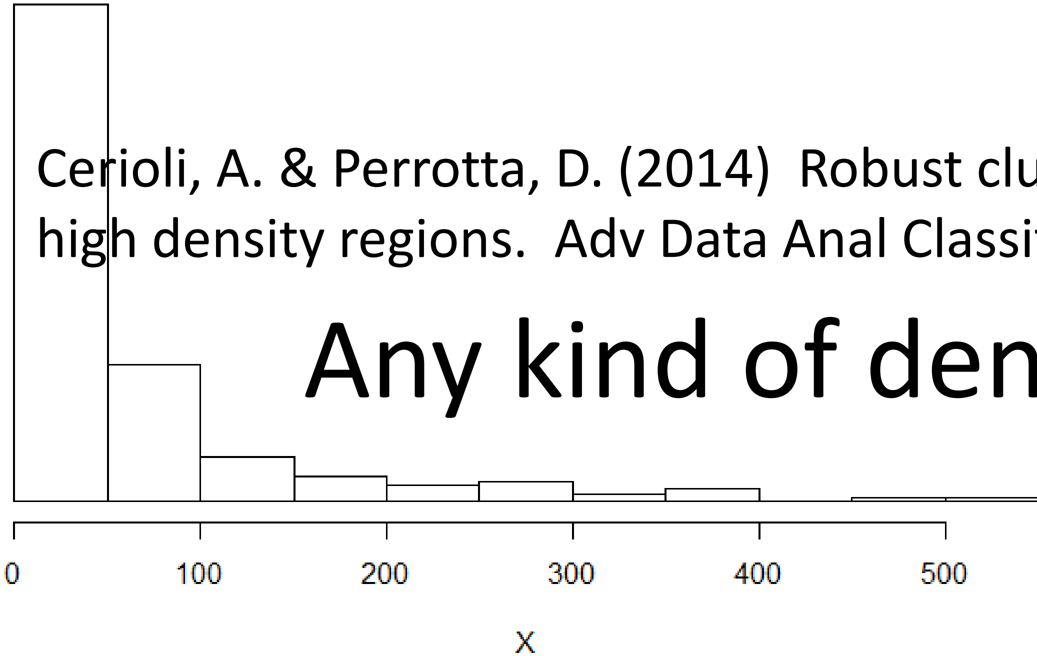


Fig. 1 Spices data set: scatter plot of value (y) and quantity (x), together with cluster membership assigned

tabase

Cerioli, A. & Perrotta, D. (2014) Robust clustering around regression lines with high density regions. Adv Data Anal Classif 8, 5-26

Any kind of denoising is needed



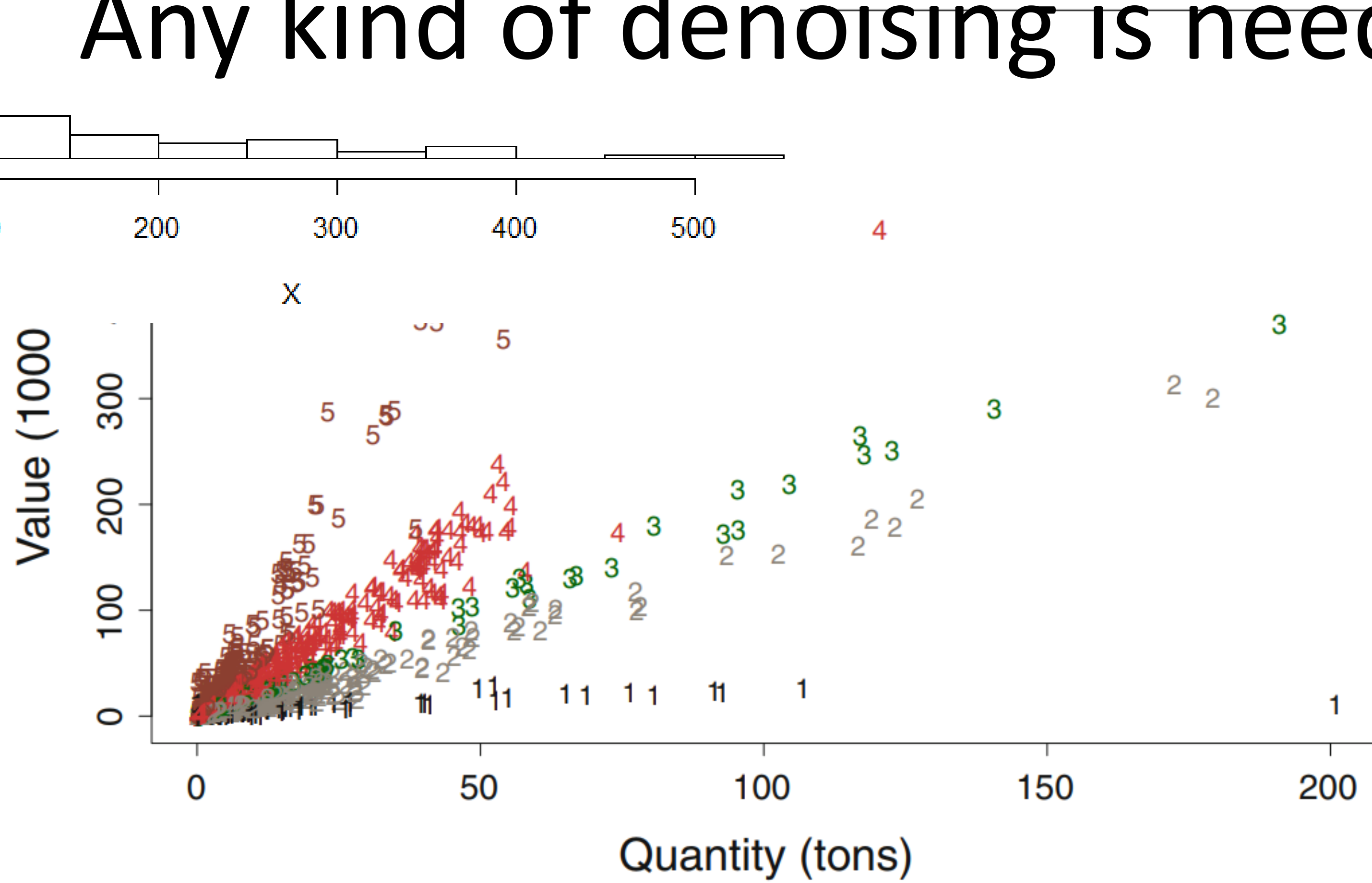


Fig. 1 Spices data set: scatter plot of value (y) and quantity (x), together with cluster membership assigned

Thinning

For avoiding the influence of concentrated contamination when estimating mixture of regressions (Cerioli and Perrotta,2014)

weighting based on density (inverse to the density)
& sampling based on this weigthing

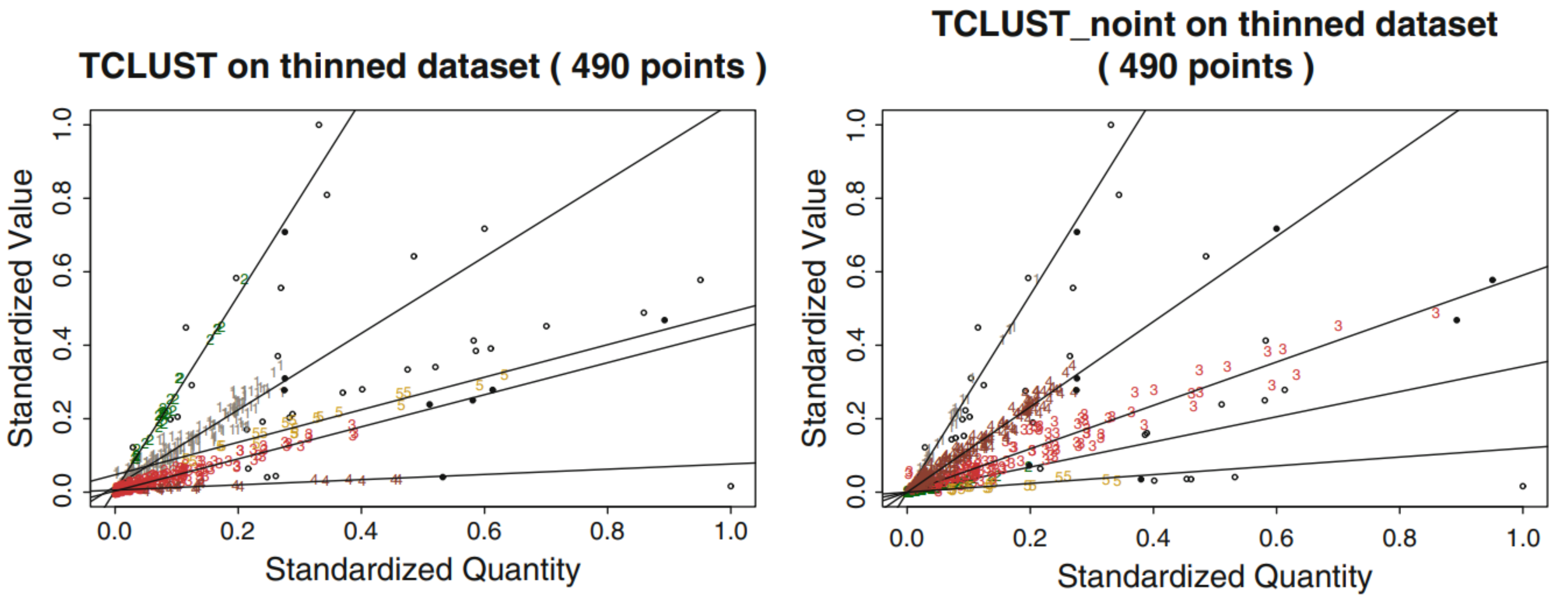


Fig. 7 Thinned Spices data set: robust fit using TCLUST-REG with $G = 5$ and trimming level 0.06. *Left panel* with intercept terms; *right panel* without intercept terms

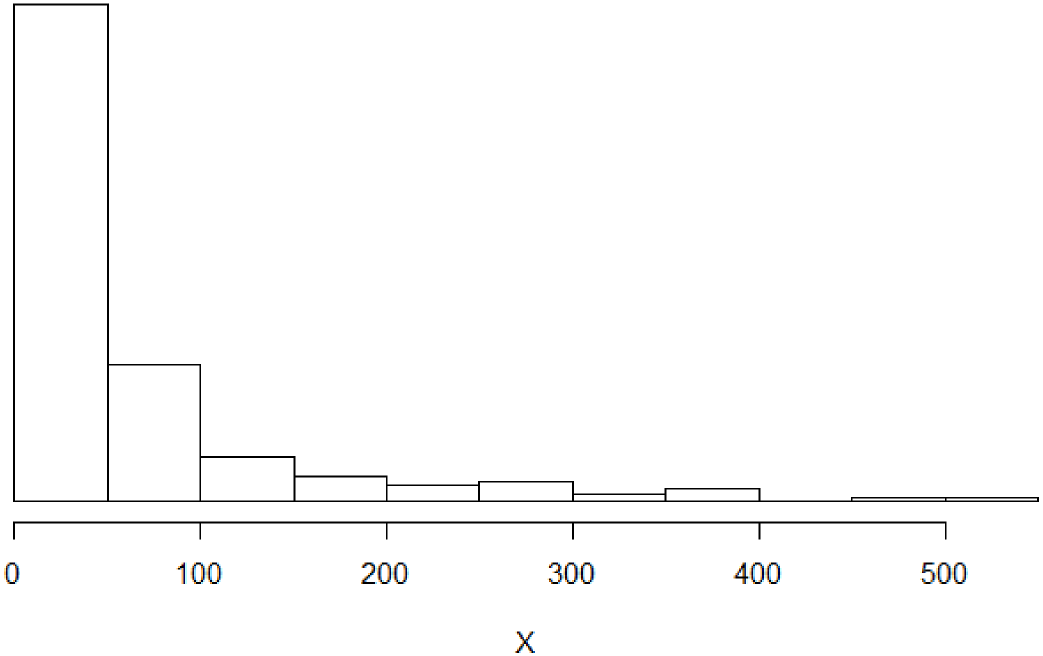
ng

ntrated contamination when

Cerioli and Perrotta,2014)

the density)

g



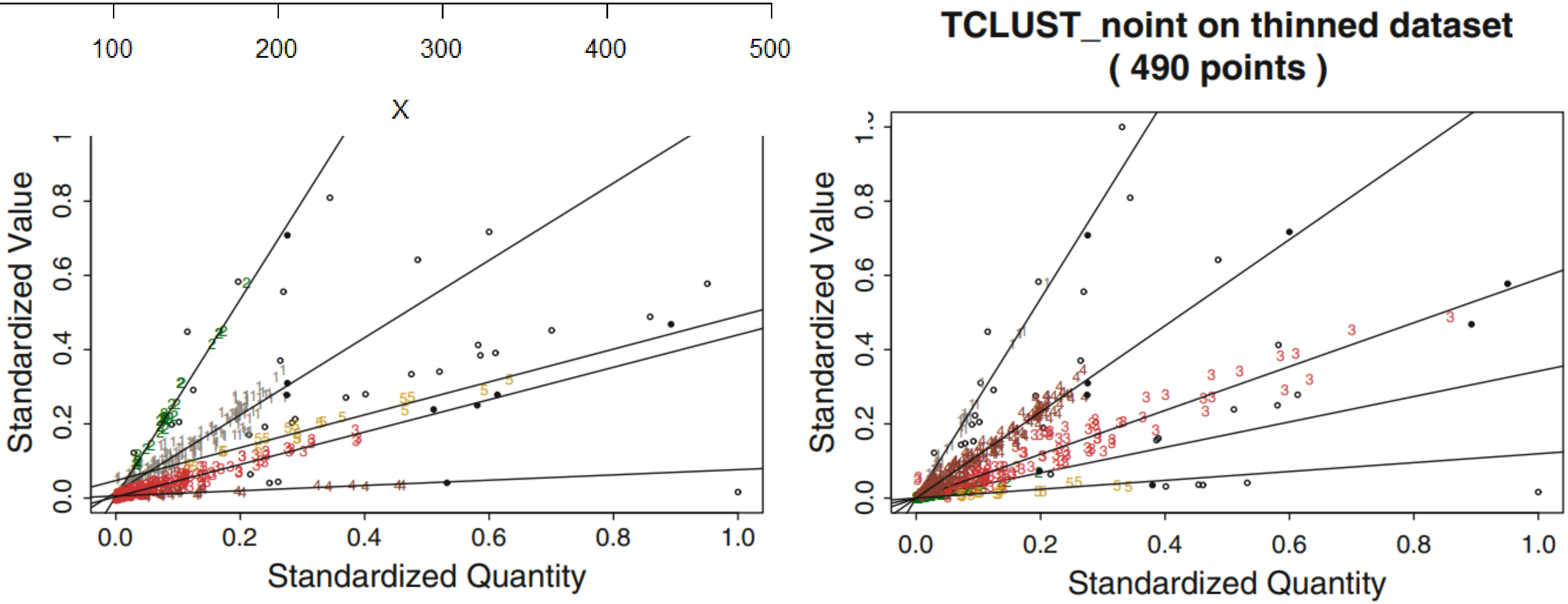
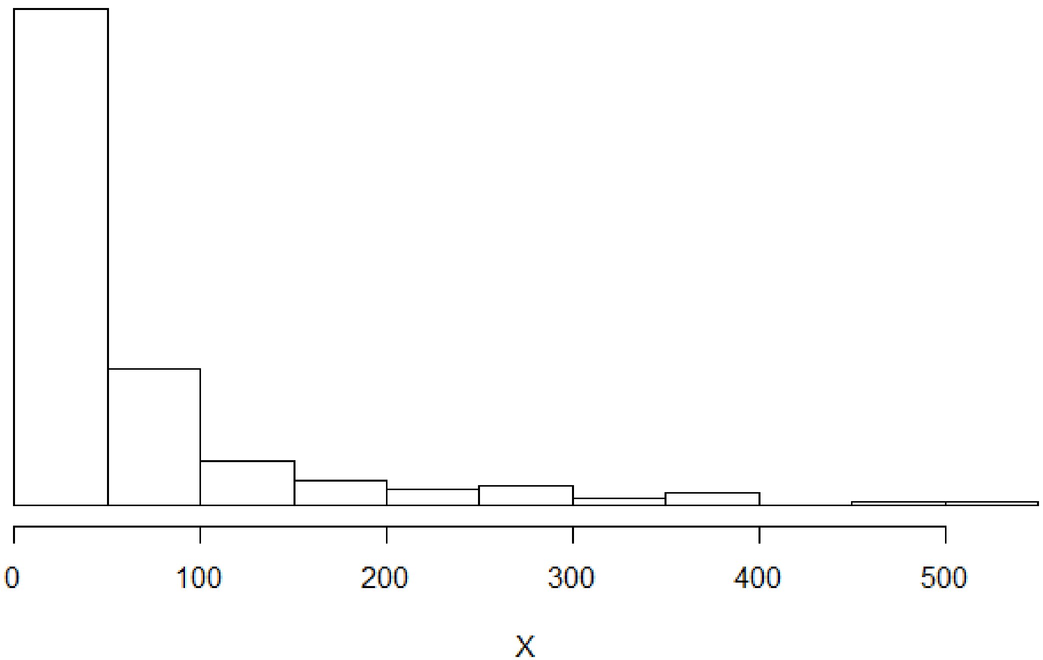


Fig. 7 Thinned Spices data set: robust fit using TCLUST-REG with $G = 5$ and trimming level 0.06. *Left panel* with intercept terms; *right panel* without intercept terms



ng

ntrated contamination when
Cerioli and Perrotta,2014)

the density)

TCLUST_noint on thinned dataset
(490 points)

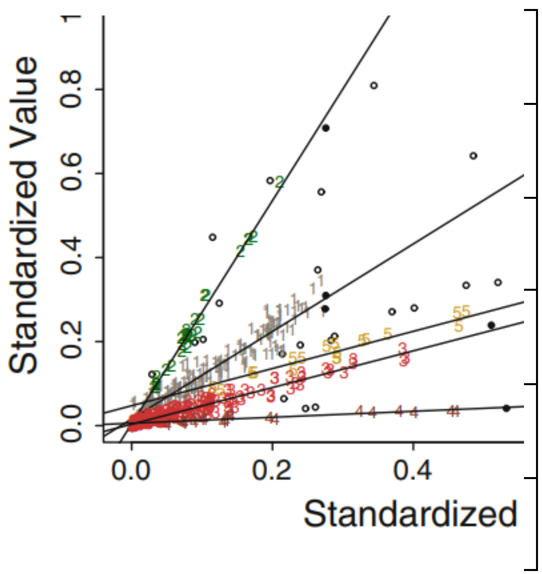
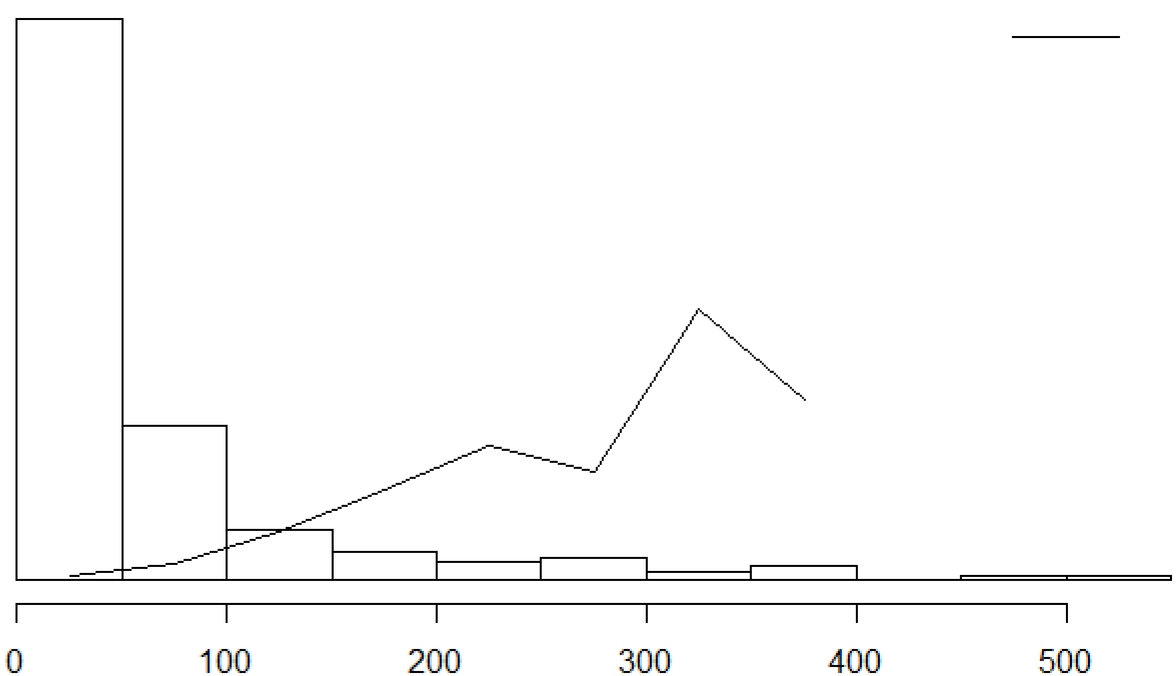


Fig. 7 Thinned Spices data set
panel with intercept terms; *righ*

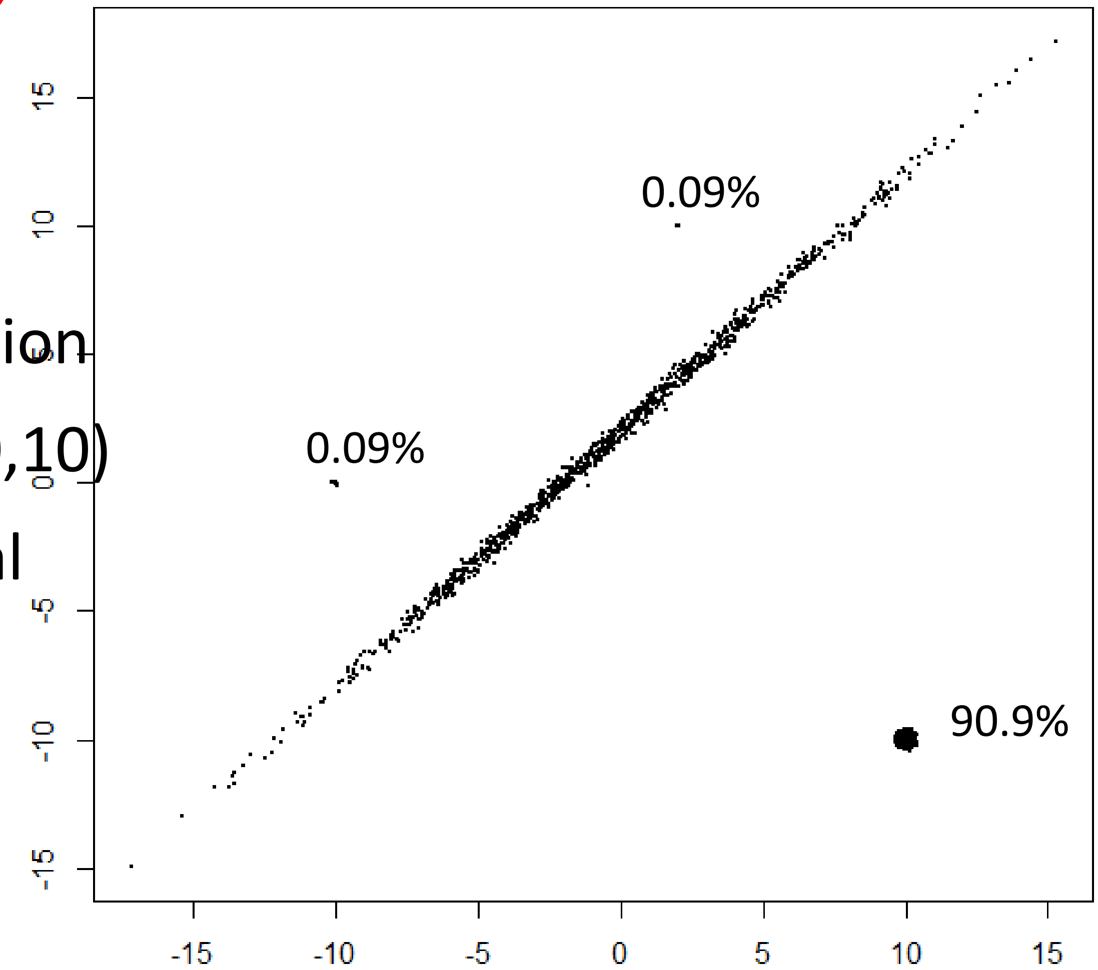


Toy example: An inusual percentage of pointwise contamination (more than 90%)

Pointwise contamination is not necessarily close to the regression line

A huge challenge, even, for Robust Statistics

Pointwise contamination (90.9%!!) located (-10,10) (0.09%) in 2 additional locations

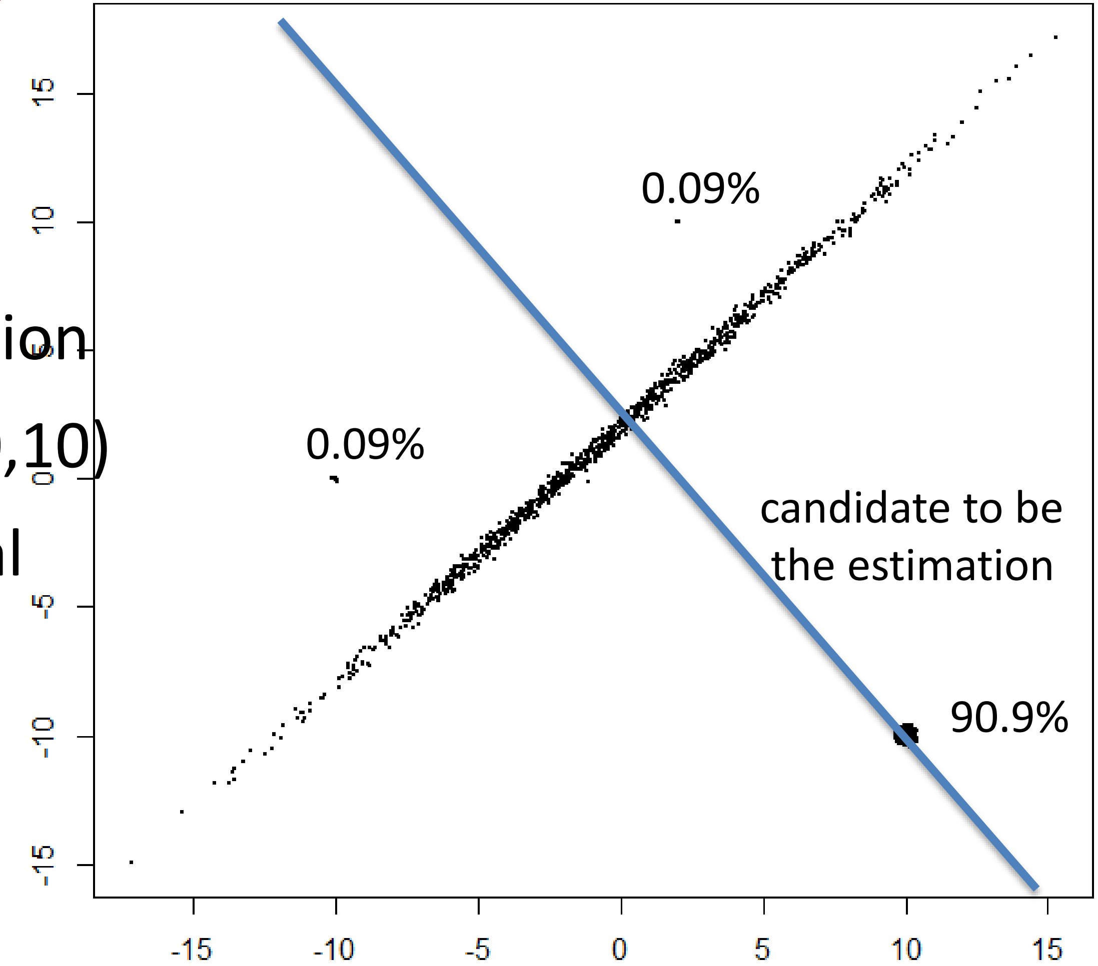


Toy example: An inusual percentage of pointwise contamination (more than 90%)

Pointwise contamination is not necessarily close to the regression line

A huge challenge, even, for Robust Statistics

Pointwise contamination (90.9%!!) located (-10,10) (0.09%) in 2 additional locations

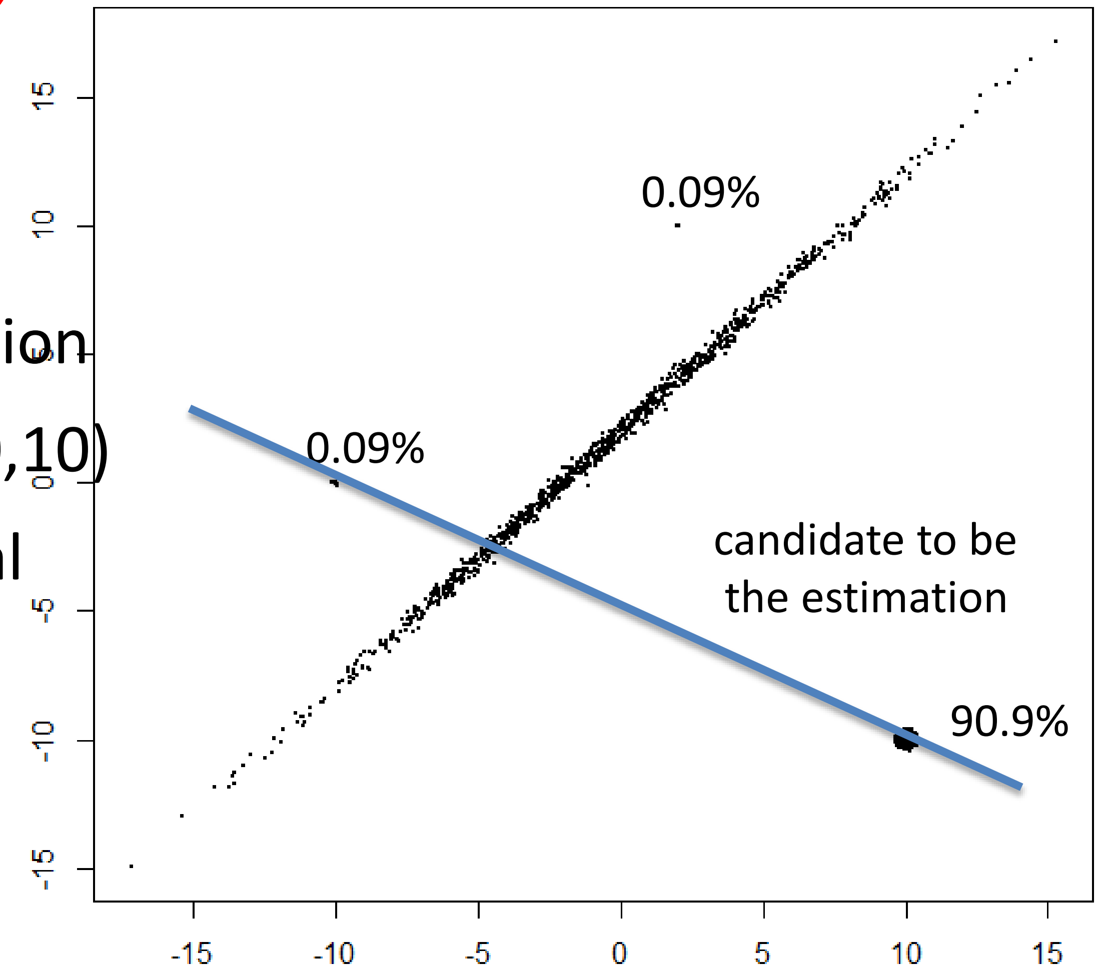


Toy example: An inusual percentage of pointwise contamination (more than 90%)

Pointwise contamination is not necessarily close to the regression line

A huge challenge, even, for Robust Statistics

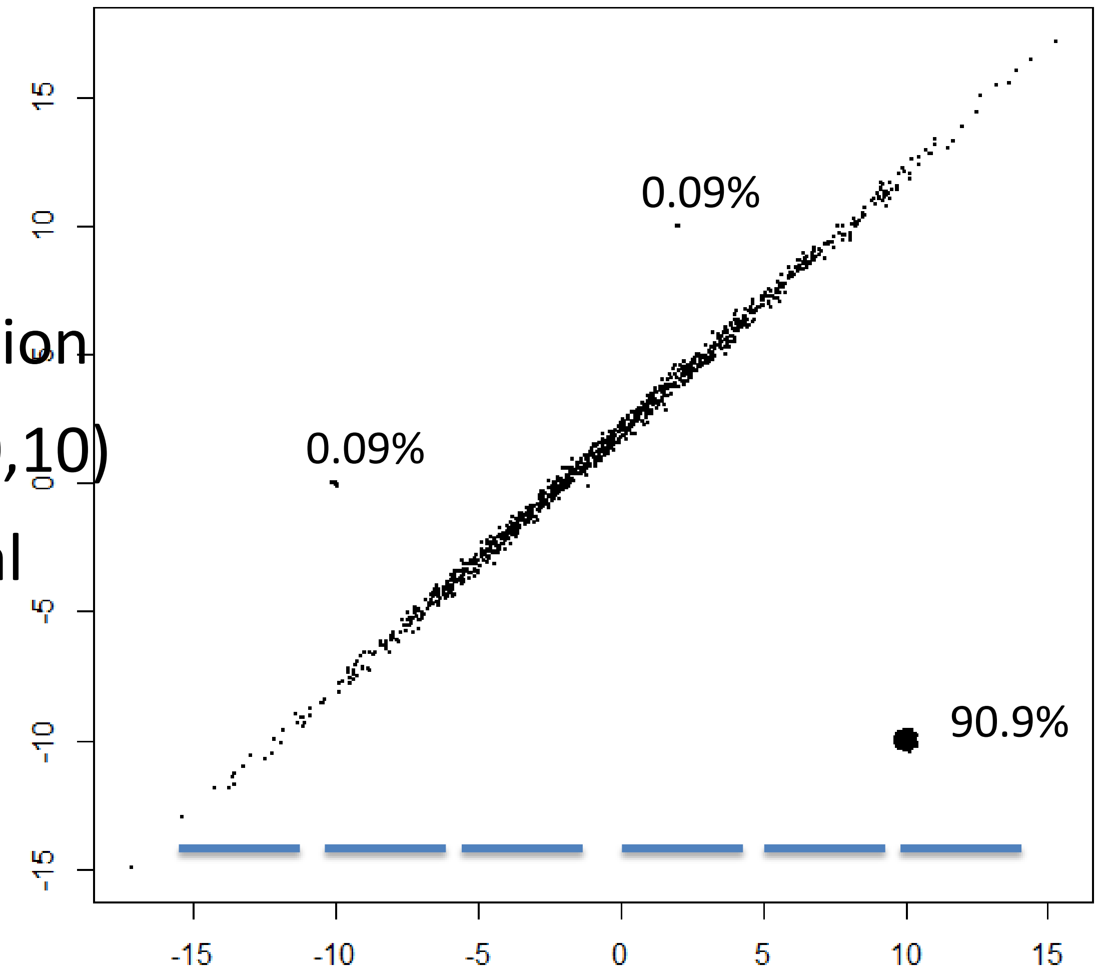
Pointwise contamination (90.9%!!) located (-10,10) (0.09%) in 2 additional locations



Toy example: An inusual percentage of pointwise contamination (more than 90%)

We are interested in a regression solution based on the whole range of x

Pointwise contamination (90.9%!!) located (-10,10) (0.09%) in 2 additional locations

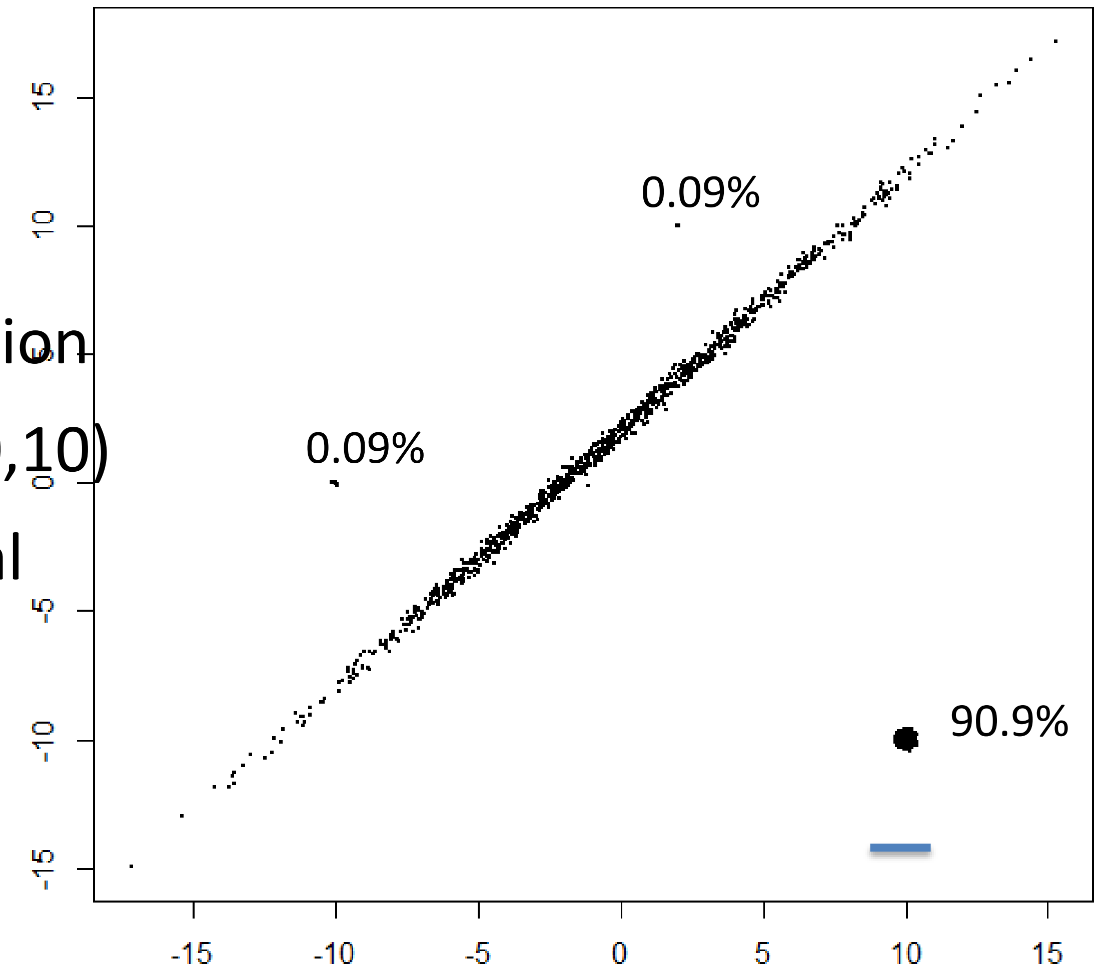


Toy example: An inusual percentage of pointwise contamination (more than 90%)

We are interested in a regression solution based on the whole range of x

Pointwise contamination (90.9%!!) located (-10,10)

(0.09%) in 2 additional locations

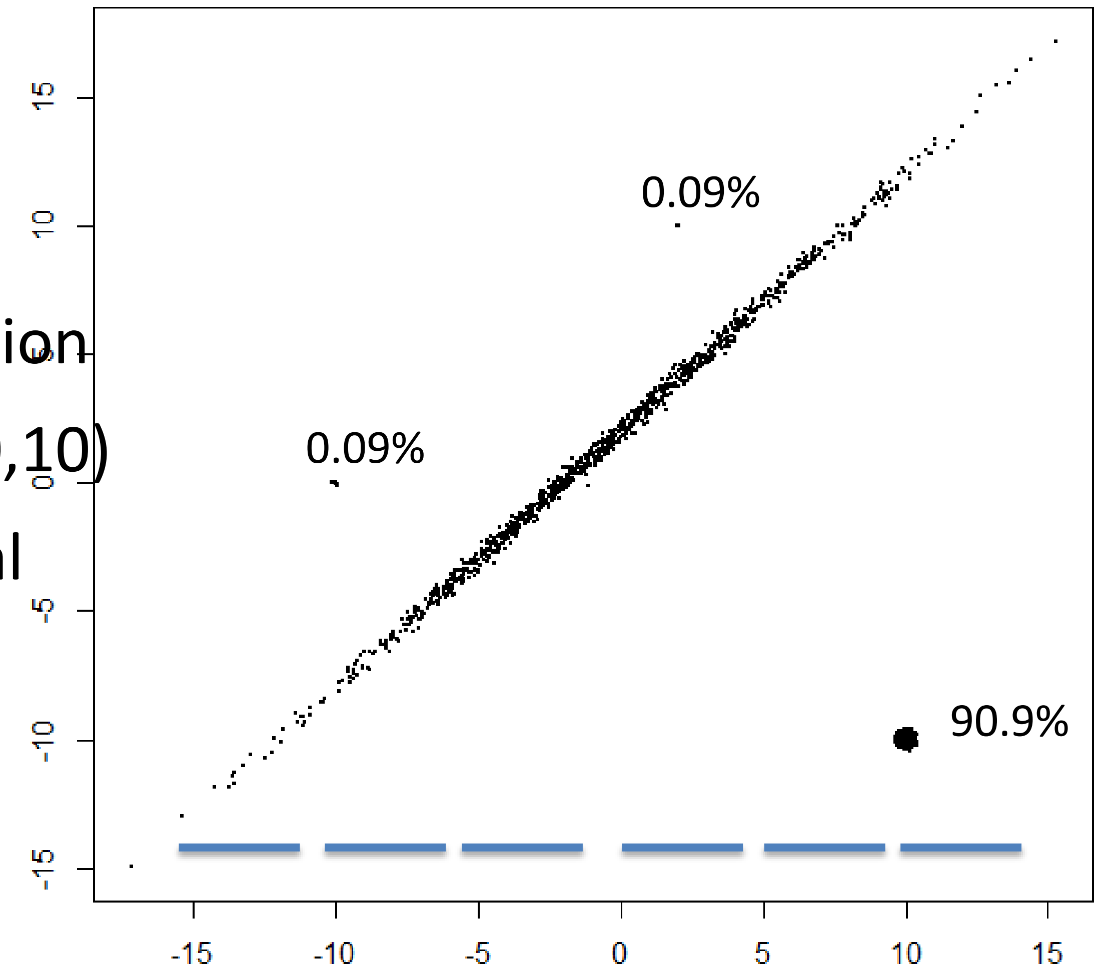


Toy example: An inusual percentage of pointwise contamination (more than 90%)

We are interested in a regression solution based on the whole range of x

Pointwise contamination (90.9%!!) located (-10,10)

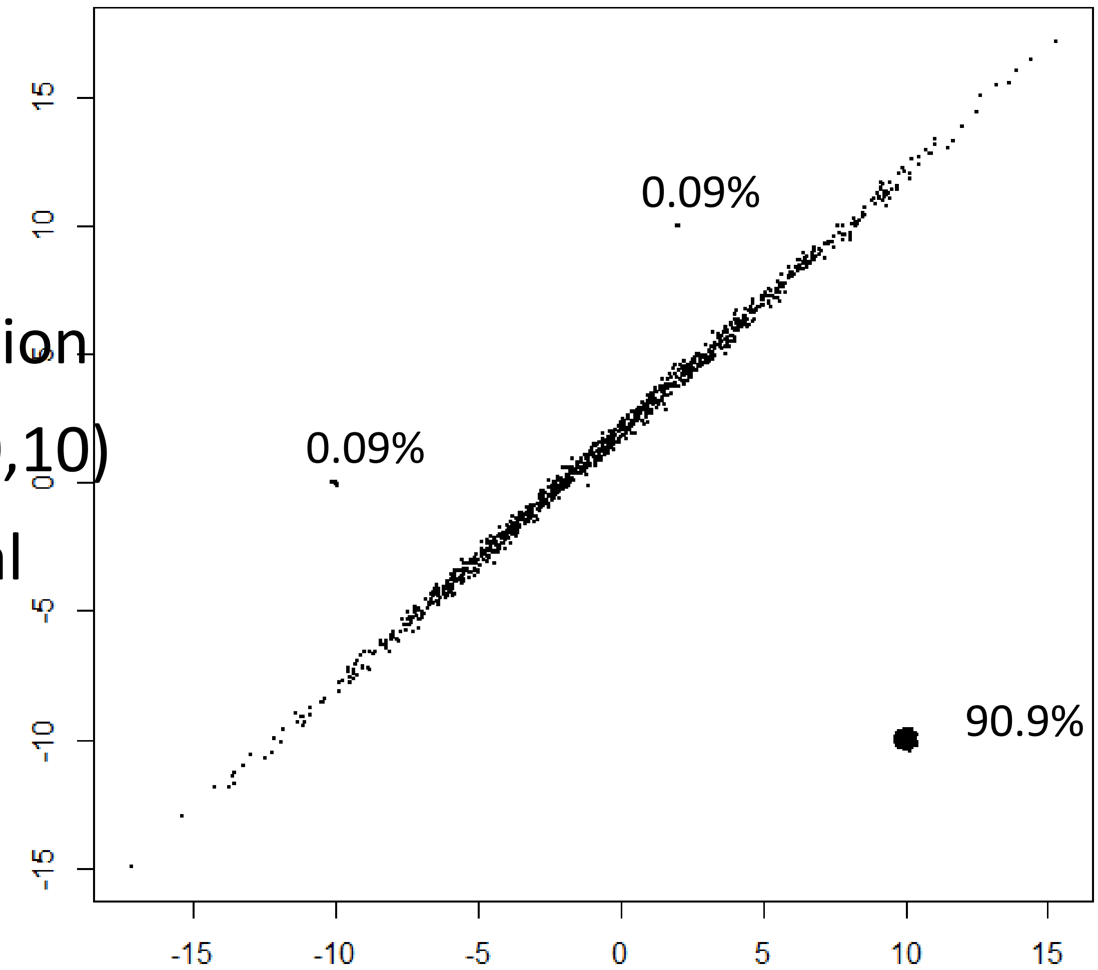
(0.09%) in 2 additional locations



Density based weighting

Thinning (Cerioli & Perrotta, 2014). De-construction of it

Pointwise contamination (90.9%!!) located (-10,10) (0.09%) in 2 additional locations

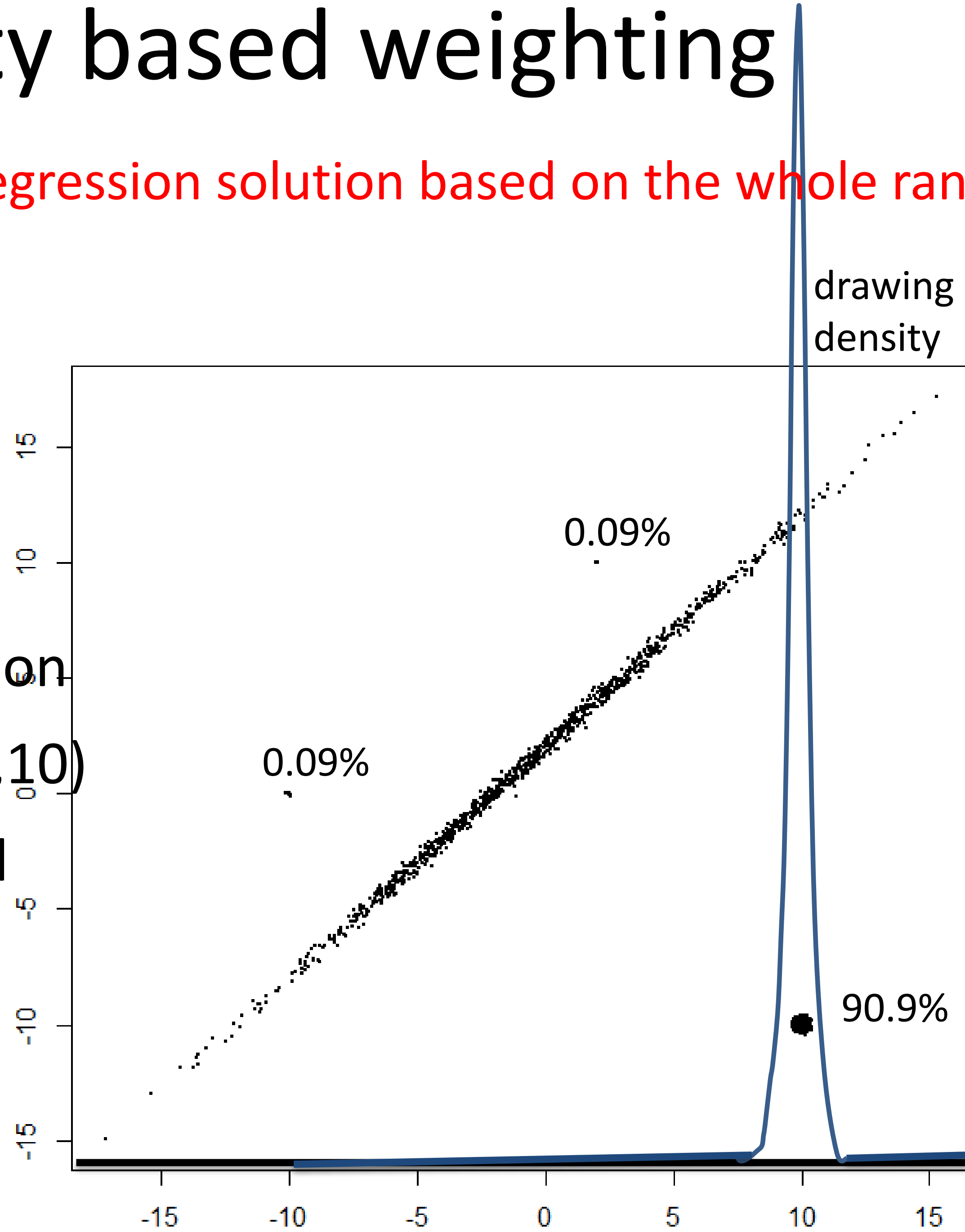


Density based weighting

We are interested in a regression solution based on the whole range of x

Pointwise contamination (90.9%!!) located (-10,10)

(0.09%) in 2 additional locations

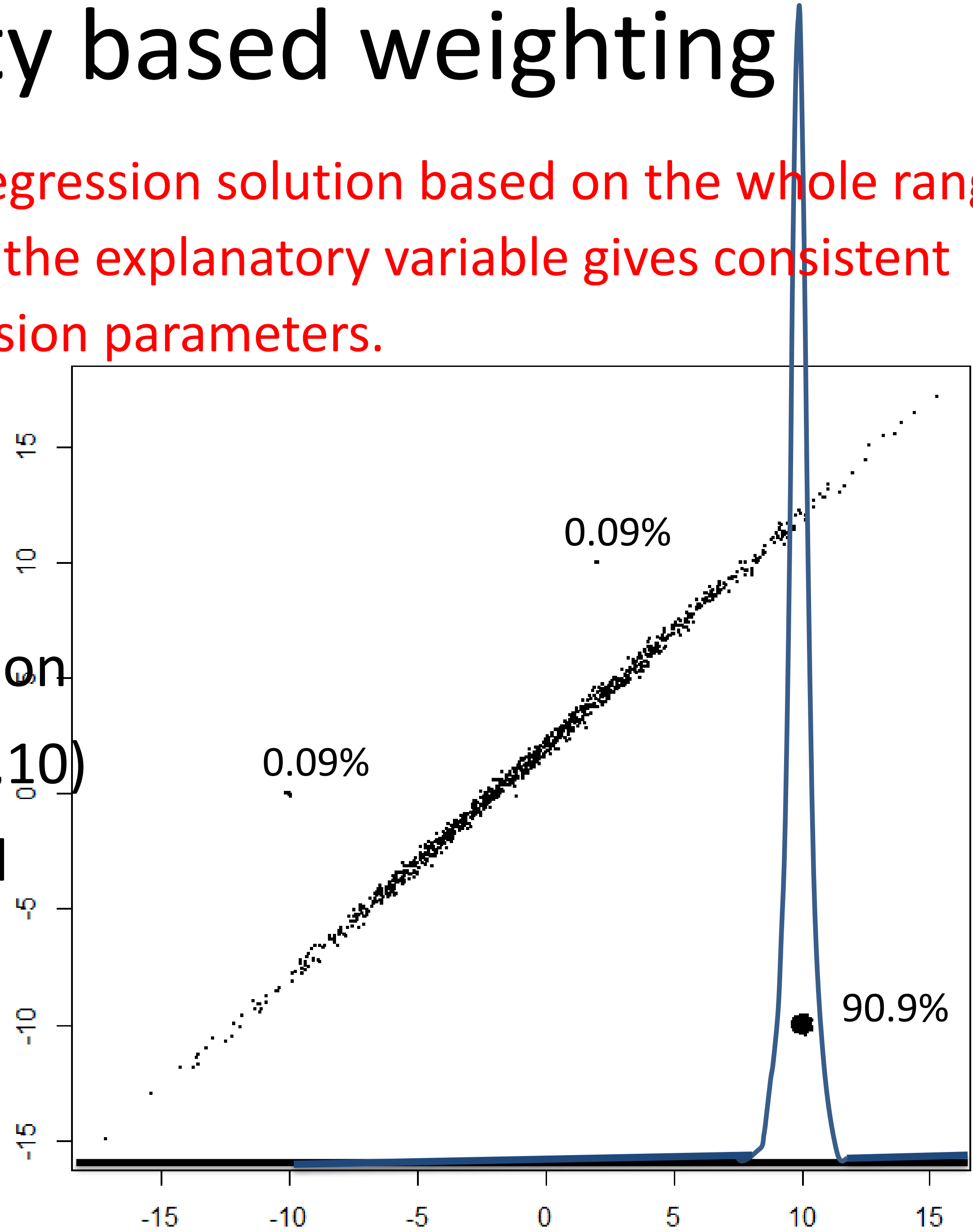


Density based weighting

We are interested in a regression solution based on the whole range of x

Any weighting based on the explanatory variable gives consistent estimation of the regression parameters.

Pointwise contamination (90.9%!!) located (-10,10) (0.09%) in 2 additional locations

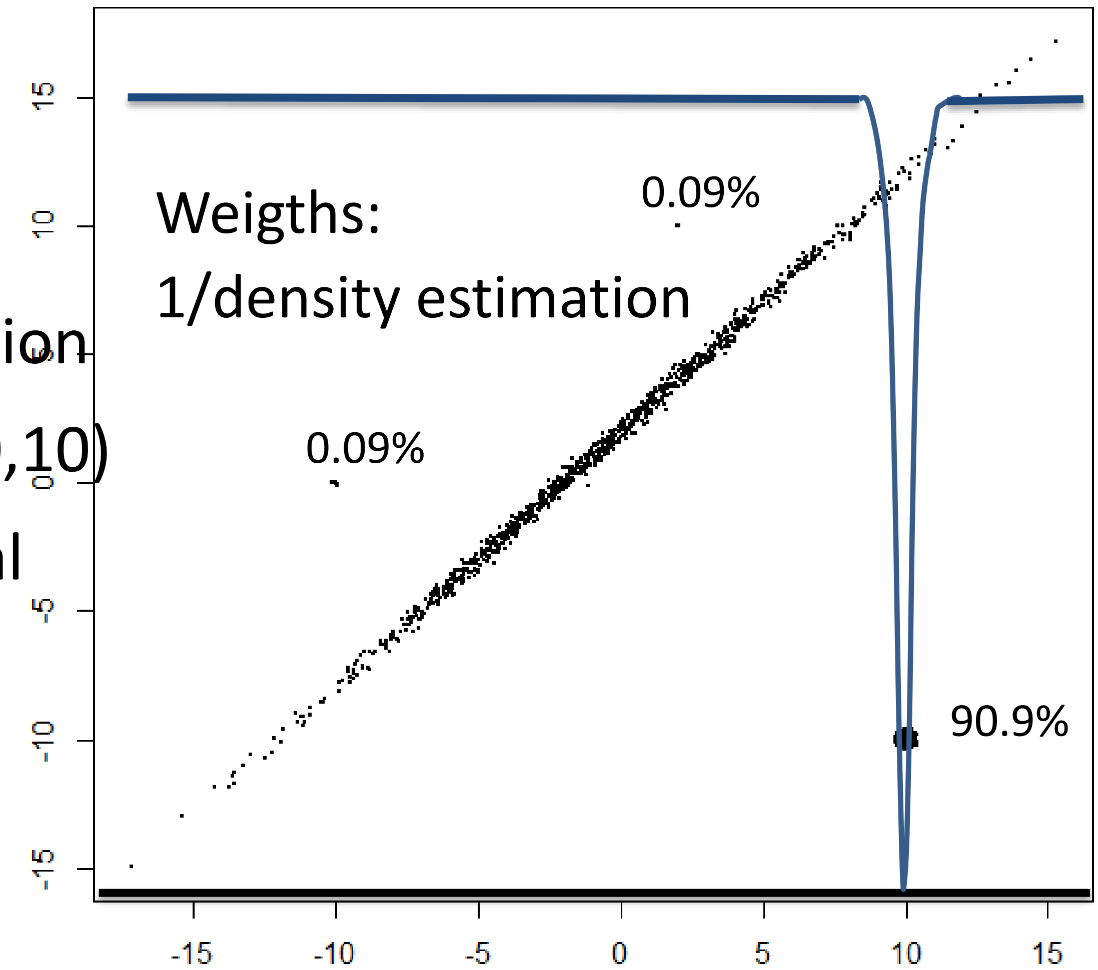


Density based weighting

Any weighting based on the explanatory variable gives consistent estimation of the regression parameters.

Pointwise contamination (90.9%!!) located (-10,10)

(0.09%) in 2 additional locations



Density based weighting

Any weighting based on the explanatory variable gives consistent estimation of the regression parameters.

To apply weighting based on density allows us to reduce the influence of concentrated contamination.

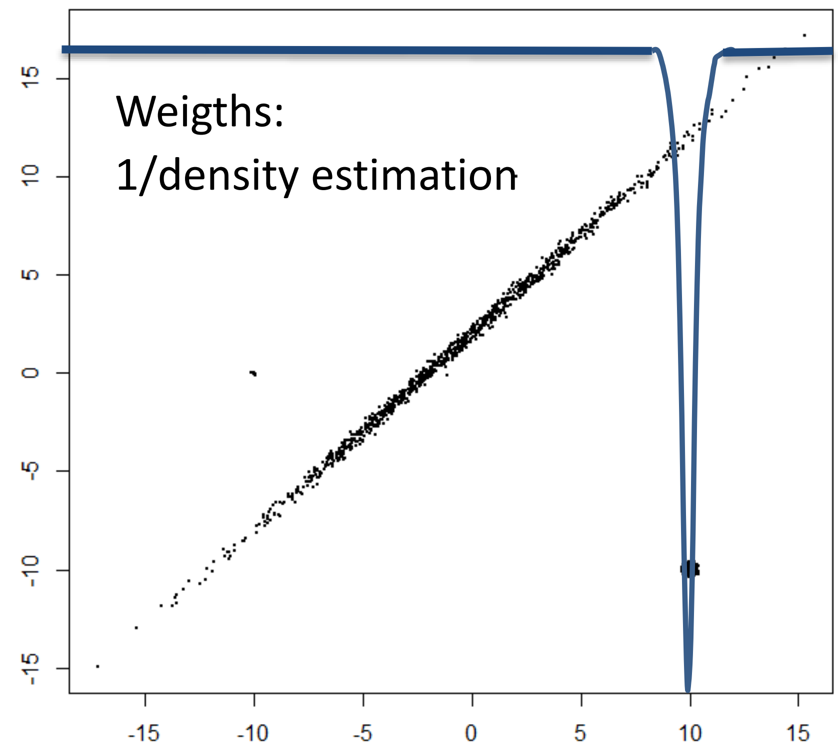
Trimming based on this weighting (not a fix proportion of observations, a fix proportion of weight)

Pointwise contamination

(90.9%!!) located (-10,10)

(0.09%) in 2 additional locations

trimming level 20%



Density based weighting

Any weighting based on the explanatory variable gives consistent estimation of the regression parameters.

To apply weighting based on density allows us to reduce the influence of concentrated contamination.

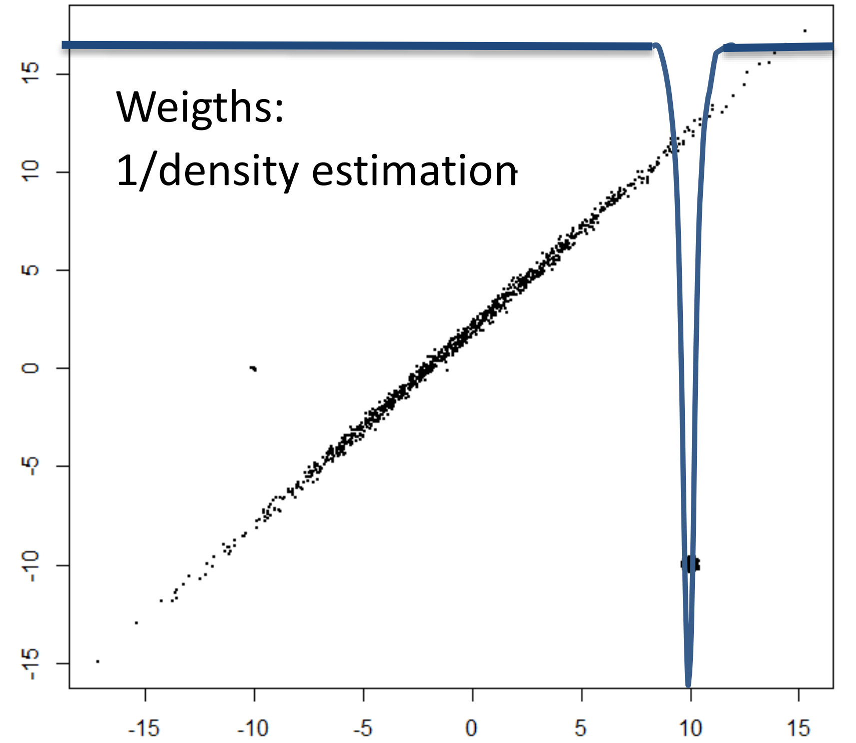
Trimming based on this weighting (not a fix proportion of observations, a fix proportion of weight)

Pointwise contamination

(90.9%!!) located (-10,10)

(0.09%) in 2 additional locations

trimming level 20%



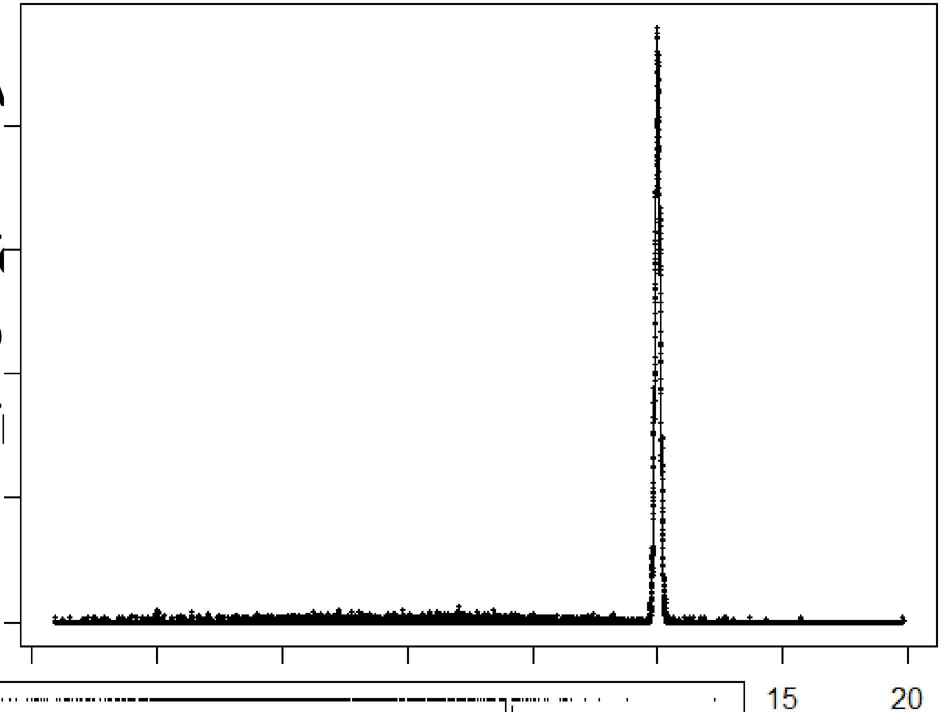
Density based weighting

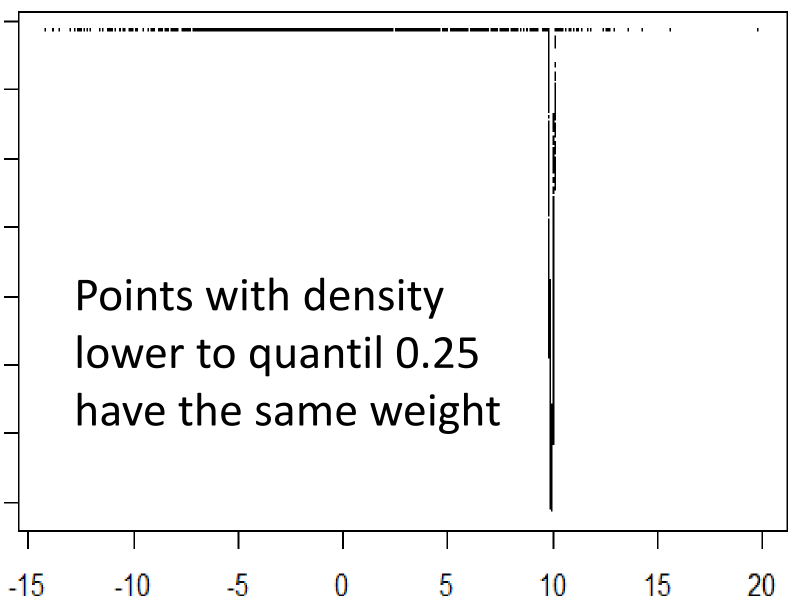
A natory variable gives consistent
e ers.

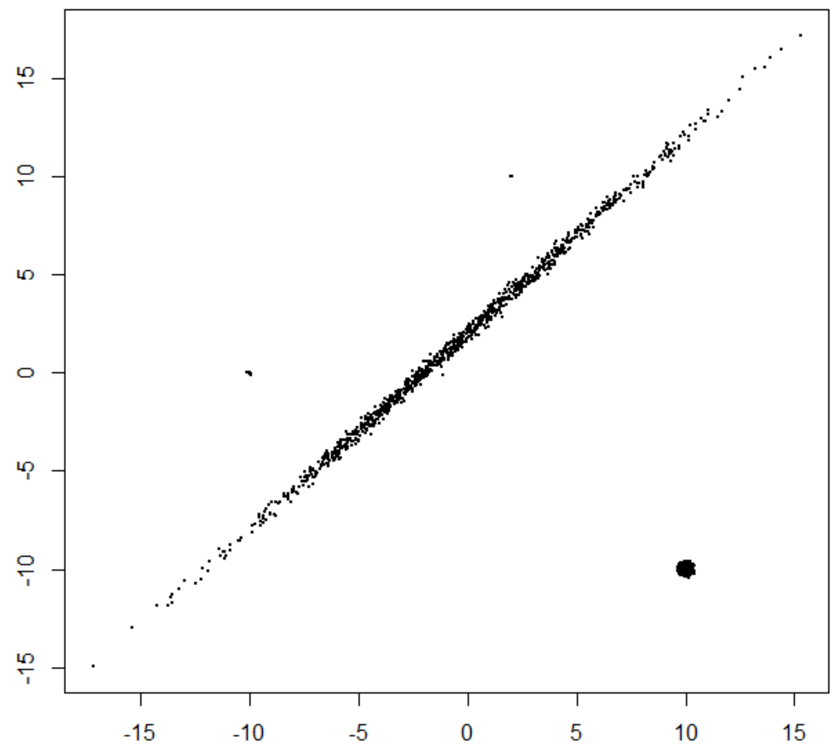
T allows us to reduce the influence
o

T t a fix proportion of observations,
a

ns







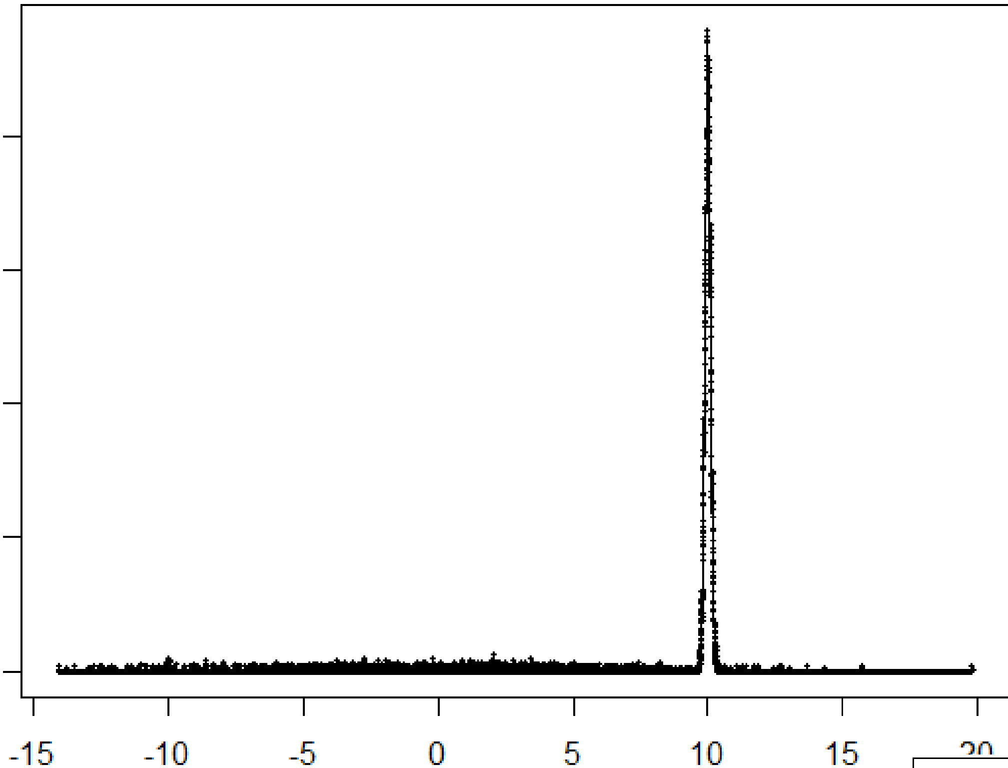
weighting

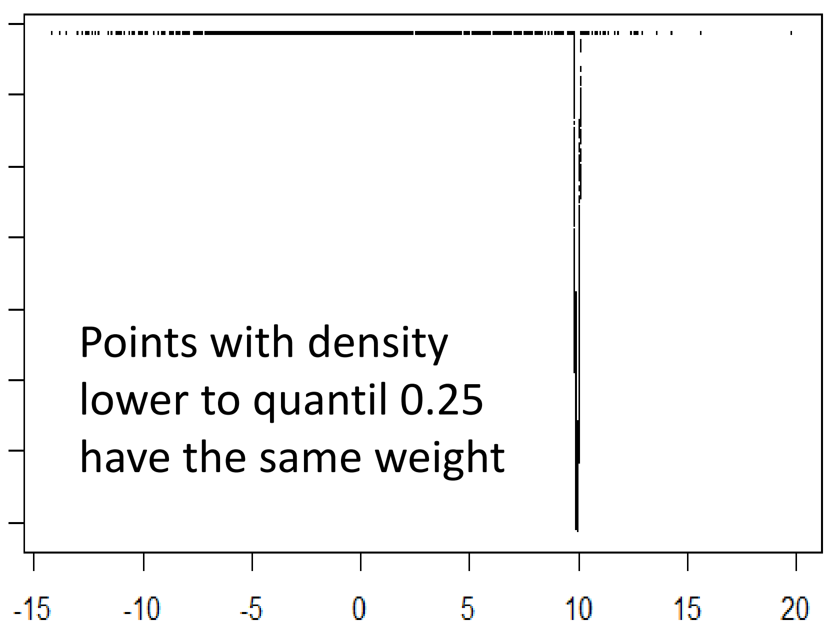
atory variable gives consistent
rs.

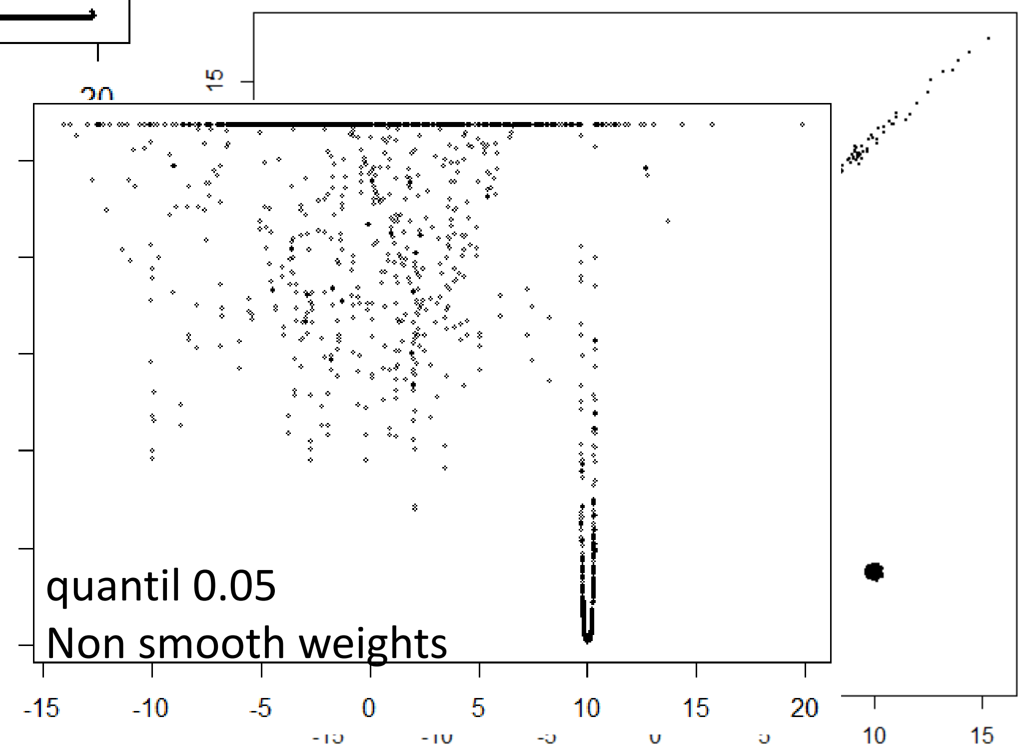
allows us to reduce the influence

a fix proportion of observations,

s







Density based weighting

Any weighting based on the explanatory variable gives consistent estimation of the regression parameters.

To apply weighting based on density allows us to reduce the influence of concentrated contamination.

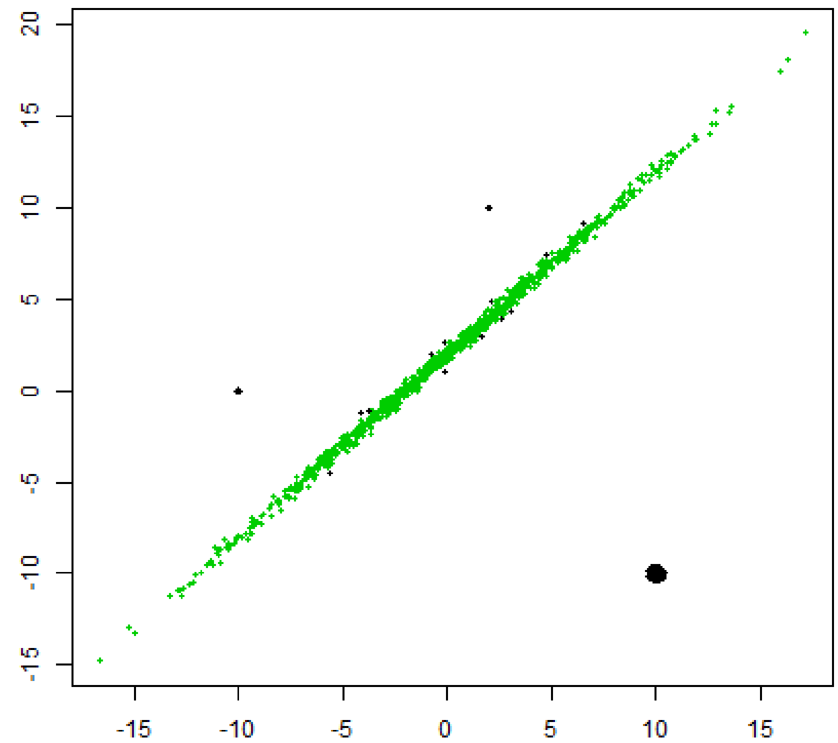
Trimming based on this weighting (not a fix proportion of observations, a fix proportion of weight)

Pointwise contamination

(90.9%!!) located (-10,10)

(0.09%) in 2 additional locations

trimming level 20%



Density based weighting

Any weighting based on the explanatory variable gives consistent estimation of the regression parameters.

To apply weighting based on density allows us to reduce the influence of concentrated contamination.

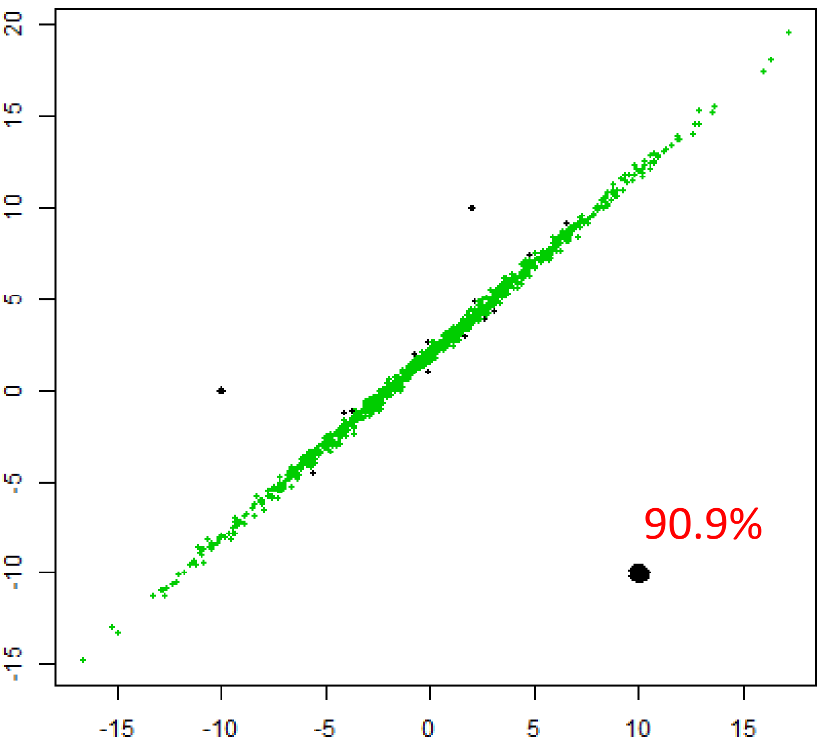
Trimming based on this weighting (not a fix proportion of observations, a fix proportion of weight)

Pointwise contamination

(90.9%!!) located (-10,10)

(0.09%) in 2 additional locations

trimming level 20%



Density based weighting

Any weighting based on the explanatory variable gives consistent estimation of the regression parameters.

To apply weighting based on density allows us to reduce the influence of concentrated contamination.

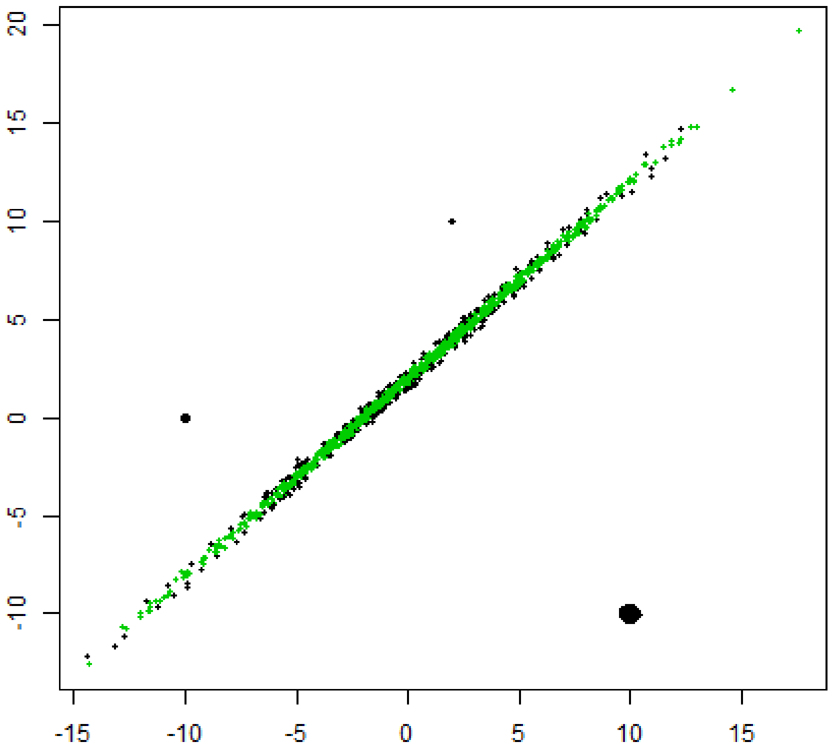
Trimming based on this weighting (not a fix proportion of observations, a fix proportion of weight)

Pointwise contamination

(90.9%!!) located (-10,10)

(0.09%) in 2 additional locations

trimming level 40%



Density based weighting

Any weighting based on the explanatory variable gives consistent estimation of the regression parameters.

To apply weighting based on density allows us to reduce the influence of concentrated contamination.

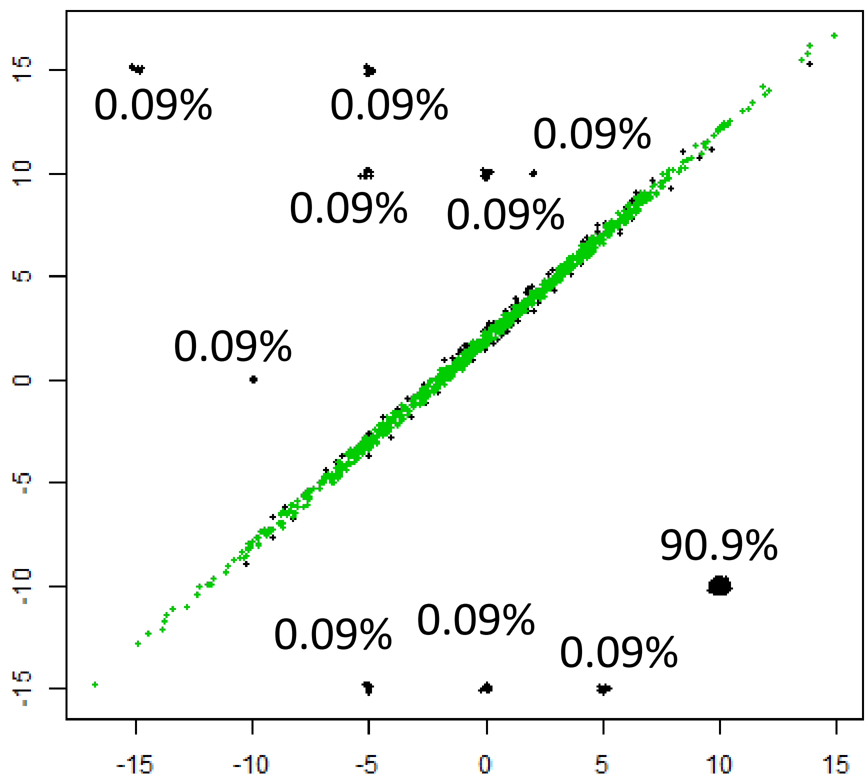
Trimming based on this weighting (not a fix proportion of observations, a fix proportion of weight)

Pointwise contamination

(90.9%!!) located (-10,10)

(0.09%) in 10 additional locations

trimming level 20%



Density based weighting (Clustering)

Any weighting based on the explanatory variable gives consistent estimation of the clustering of regression parameters.

To apply weighting based on density allows us to reduce the influence of concentrated contamination.

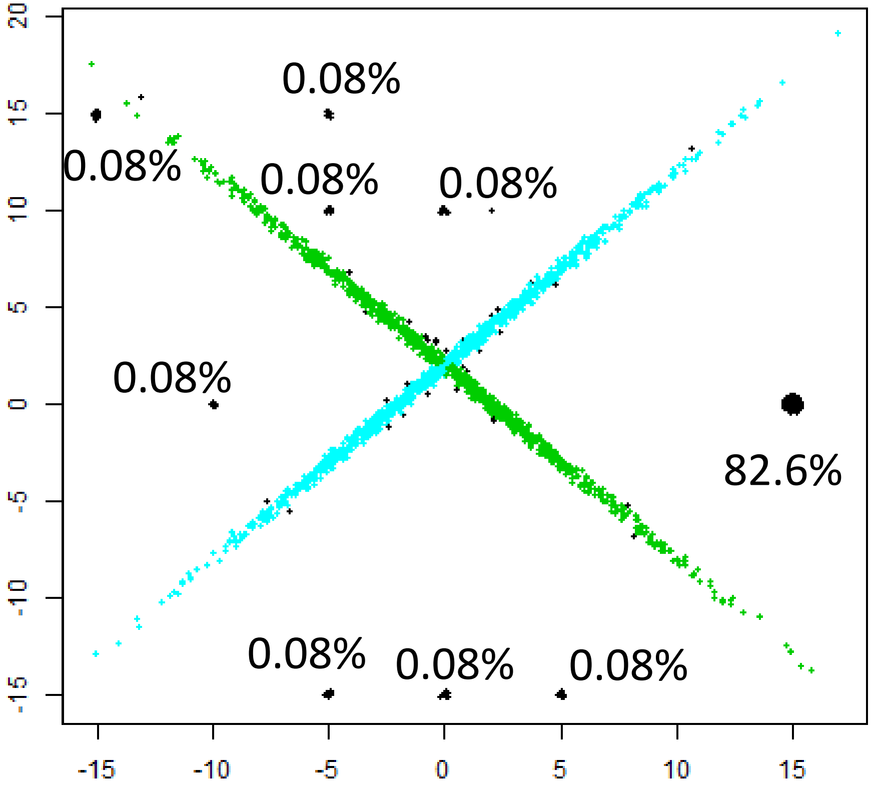
Trimming based on this weighting (not a fix proportion of observations, a fix proportion of weight)

Pointwise contamination

(82.6%!!) located (-10,10)

(0.08%) in 10 additional locations

trimming level 30%



Density based weighting (Clustering)

Any weighting based on the explanatory variable gives consistent estimation of the clustering of regression parameters.

To apply weighting based on density allows us to reduce the influence of concentrated contamination.

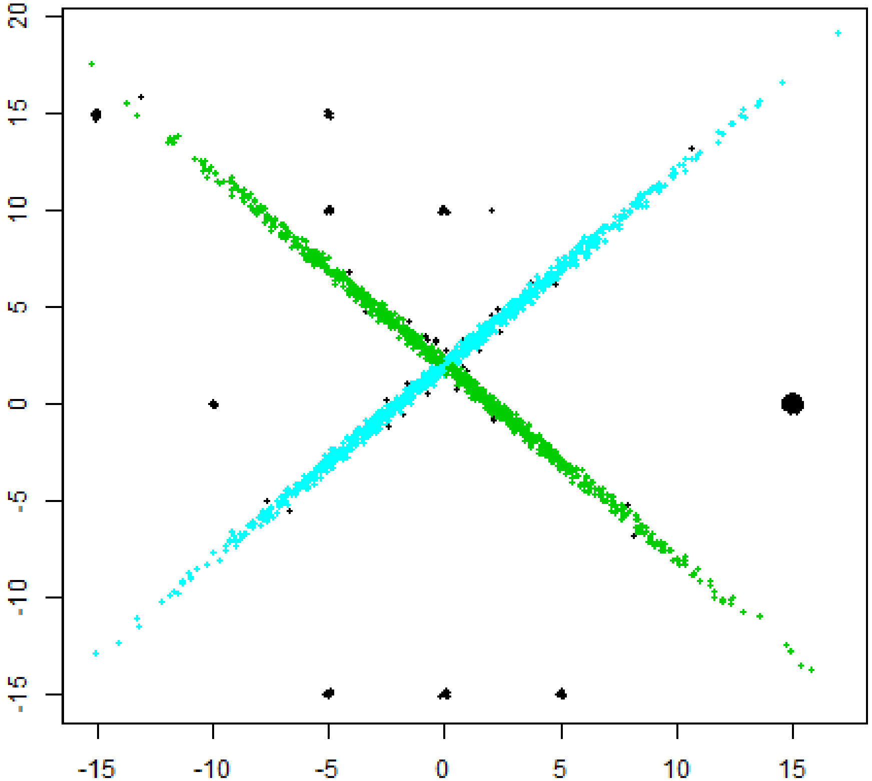
Trimming based on this weighting (not a fix proportion of observations, a fix proportion of weight)

Pointwise contamination

(82.6%!!) located (-10,10)

(0.08%) in 10 additional locations

trimming level 30%



Density based weighting

Any weighting based on the explanatory variable gives consistent estimation of the clustering of regression parameters.

To apply weighting based on density allows us to reduce the influence of concentrated contamination.

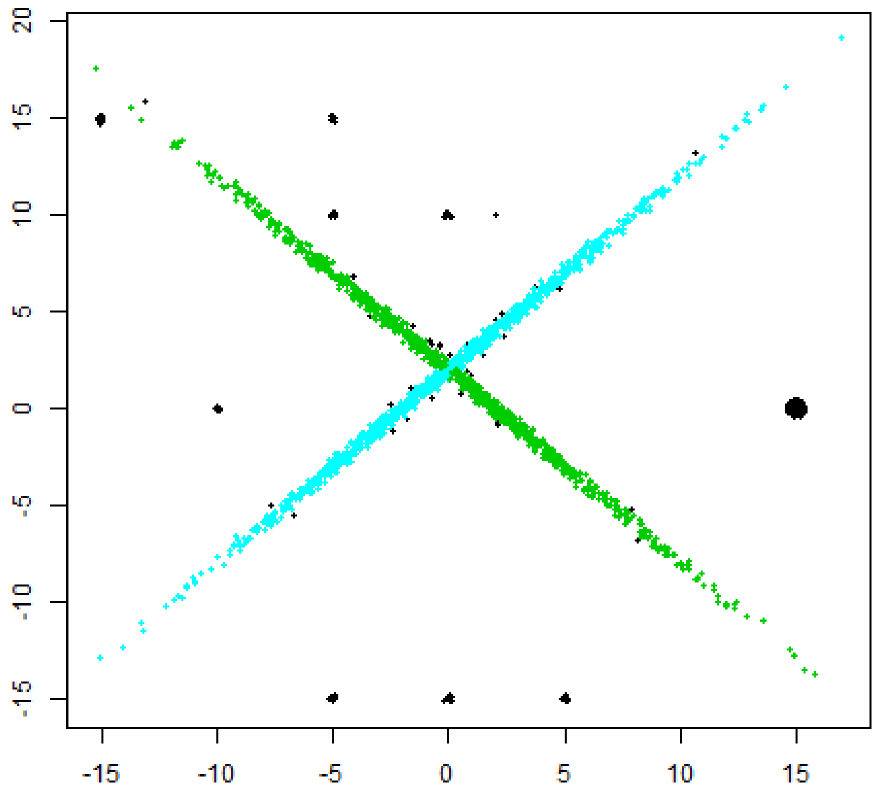
Trimming based on this weighting (not a fix proportion of observations, a fix proportion of weight)

Pointwise contamination

(82.6%!!) located (-10,10)

(0.08%) in 10 additional locations

trimming level 30%



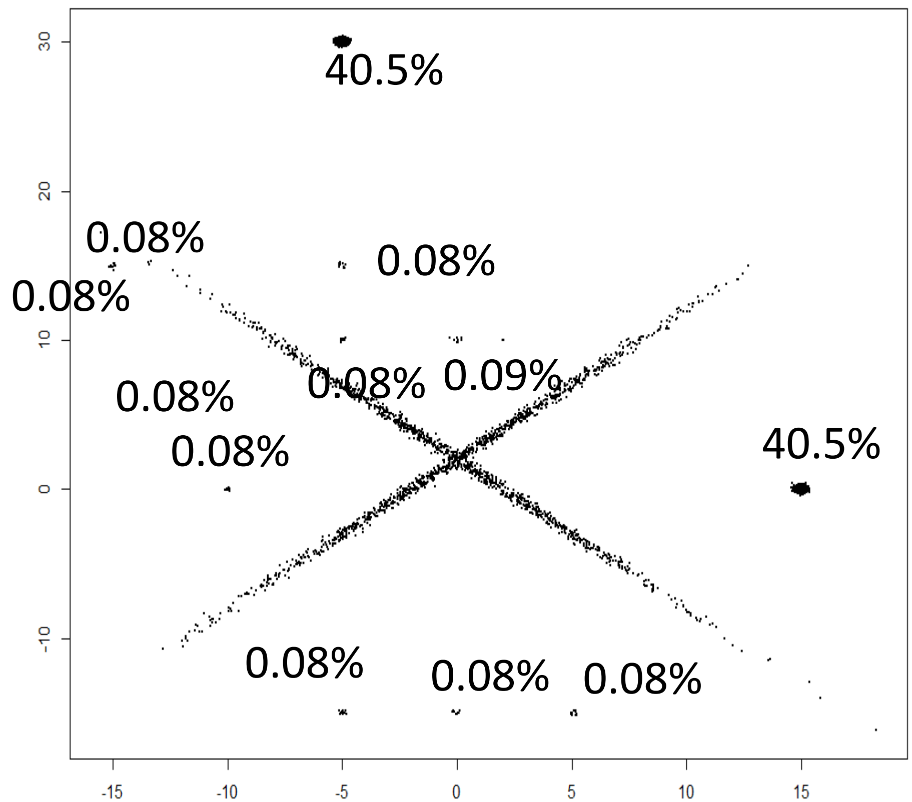
Orthogonal regression - Density based weighting

Pointwise contamination

(40.5%!!) located (0,15)

(40.5%!!) located (-5,30)

(0.08%) in 10 additional locations



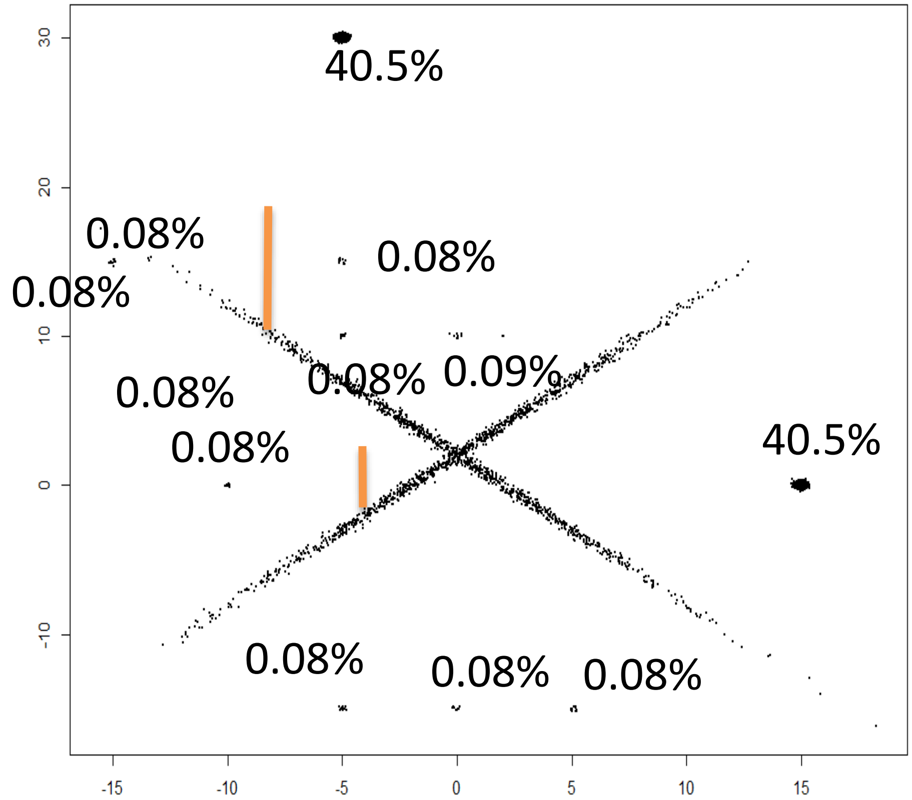
Orthogonal regression - Density based weighting

Pointwise contamination

(40.5%!!) located (0,15)

(40.5%!!) located (-5,30)

(0.08%) in 10 additional locations



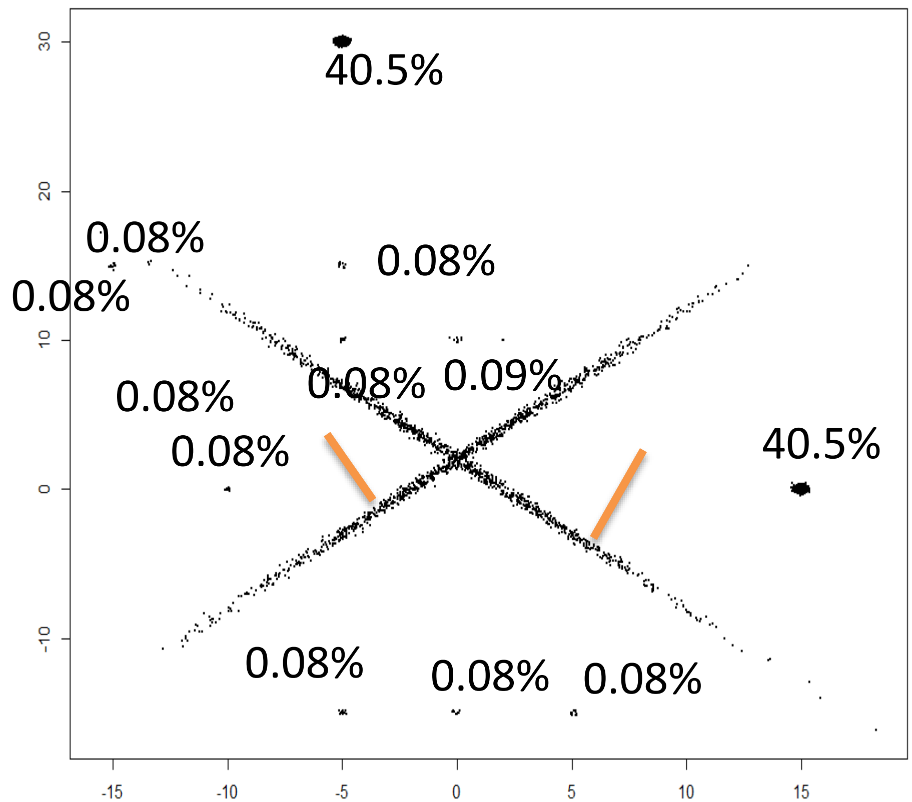
Orthogonal regression - Density based weighting

Pointwise contamination

(40.5%!!) located (0,15)

(40.5%!!) located (-5,30)

(0.08%) in 10 additional locations



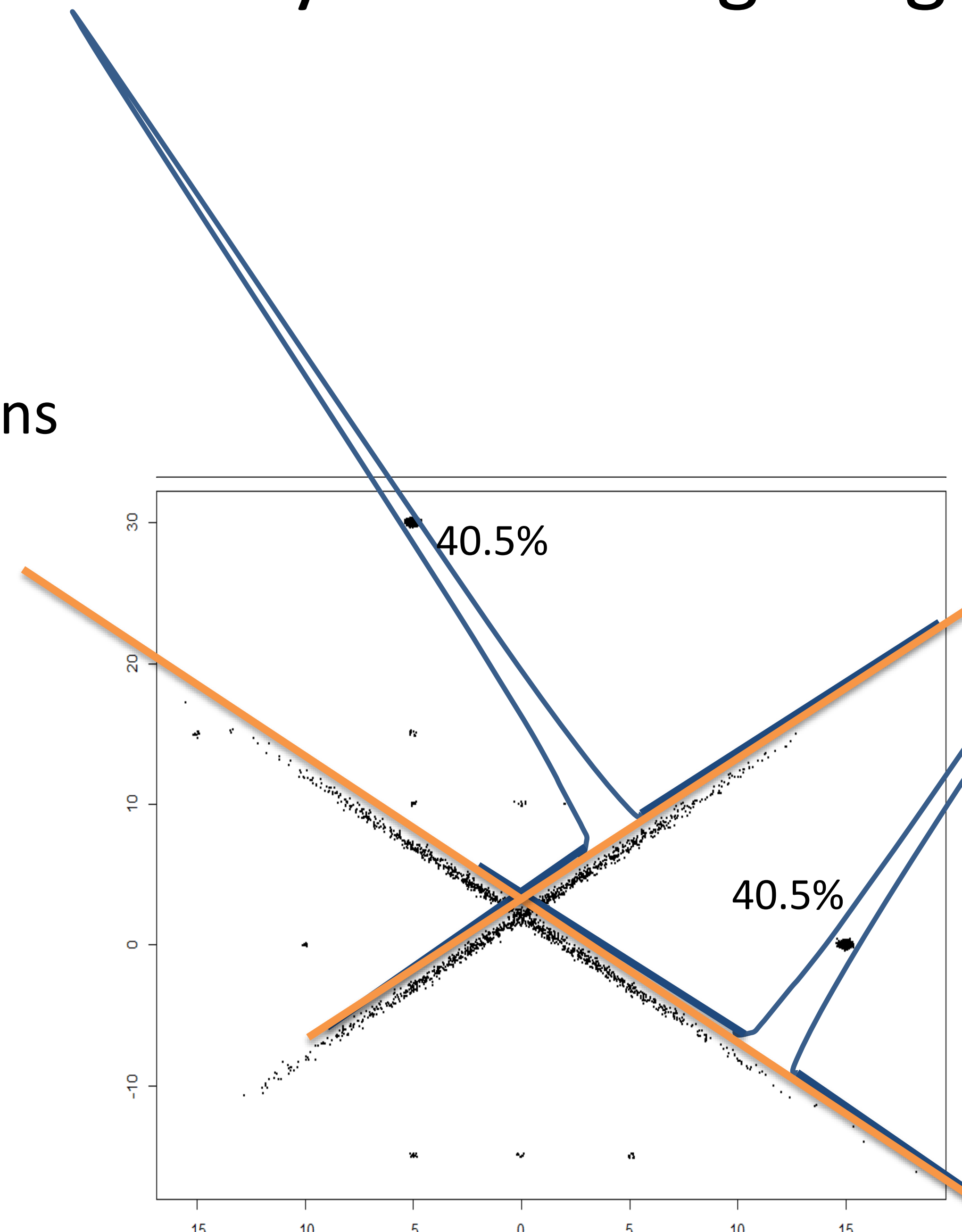
Orthogonal regression - Density based weighting

Pointwise contamination

(40.5%!!) located (0,15)

(40.5%!!) located (-5,30)

(0.08%) in 10 additional locations



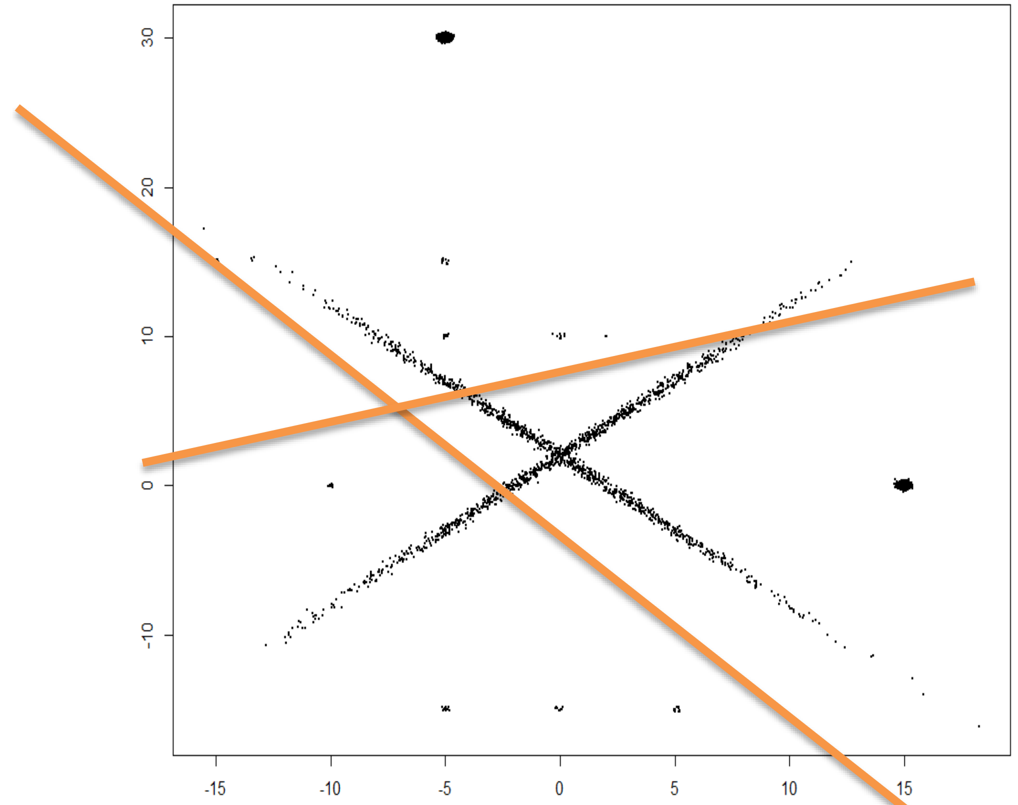
Orthogonal regression - Density based weighting

Pointwise contamination

(40.5%!!) located (0,15)

(40.5%!!) located (-5,30)

(0.08%) in 10 additional locations



Orthogonal regression - Density based weighting

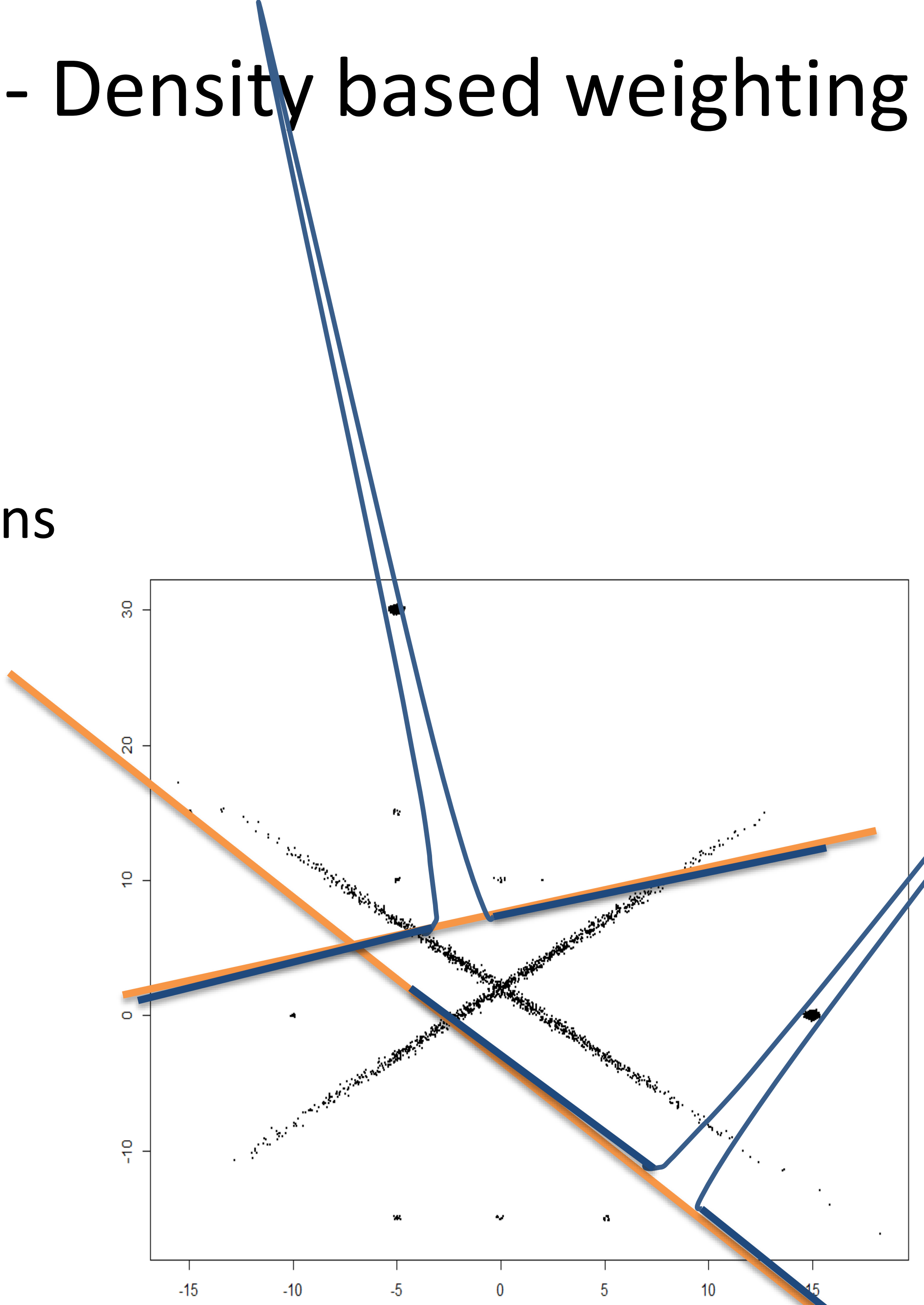
Pointwise contamination

(40.5%!!) located (0,15)

(40.5%!!) located (-5,30)

(0.08%) in 10 additional locations

It is necessary to include the density estimation in each step of the algorithm



Orthogonal regression - Density based weighting

Start with *k* random linear models

Iterations

*Assign each observation to the closest model

Estimate density in each model separately

Weight based on density

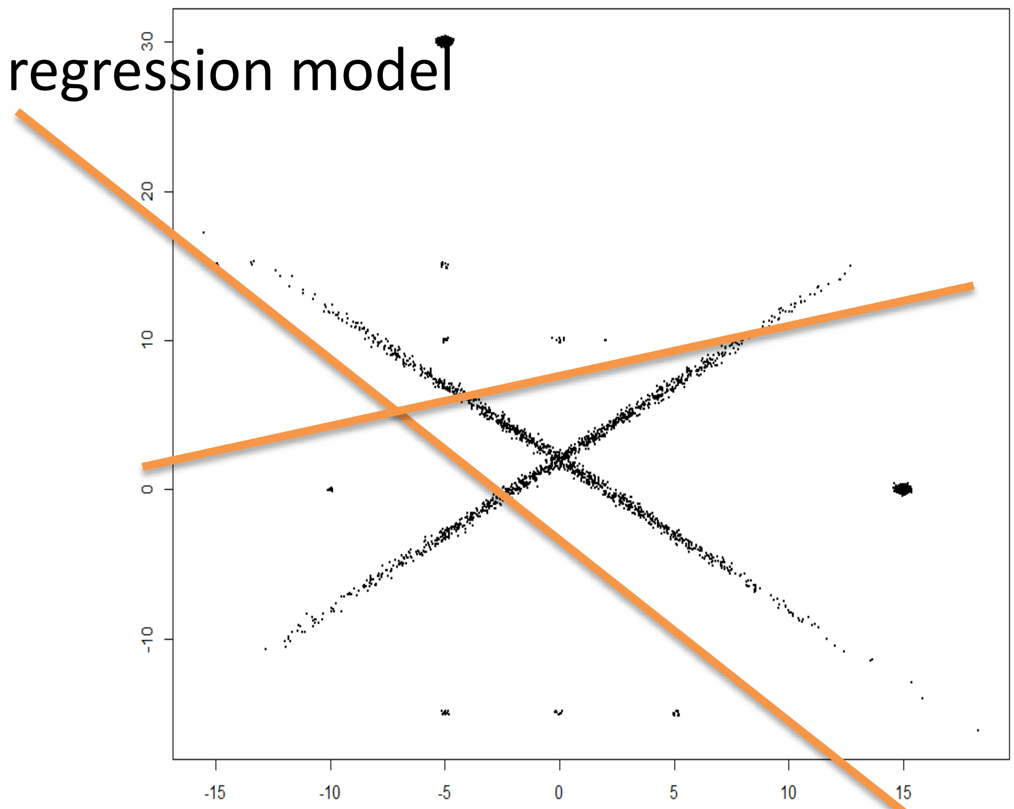
*estimation of each orthogonal regression model

Pointwise contamination

(40.5%!!) located (0,15)

(40.5%!!) located (-5,30)

(0.09%) in 10 additional locations



Orthogonal regression - Density based weighting

Start with *k* random linear models

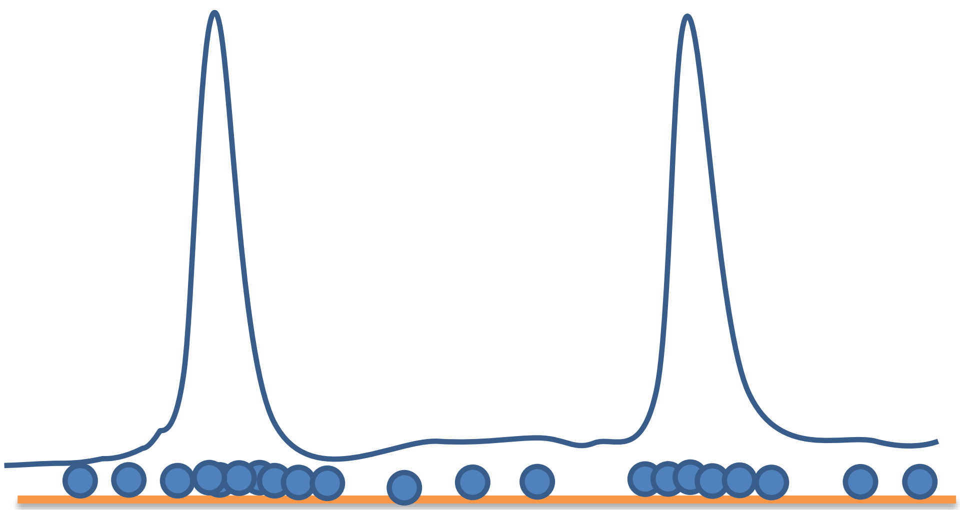
Iterations

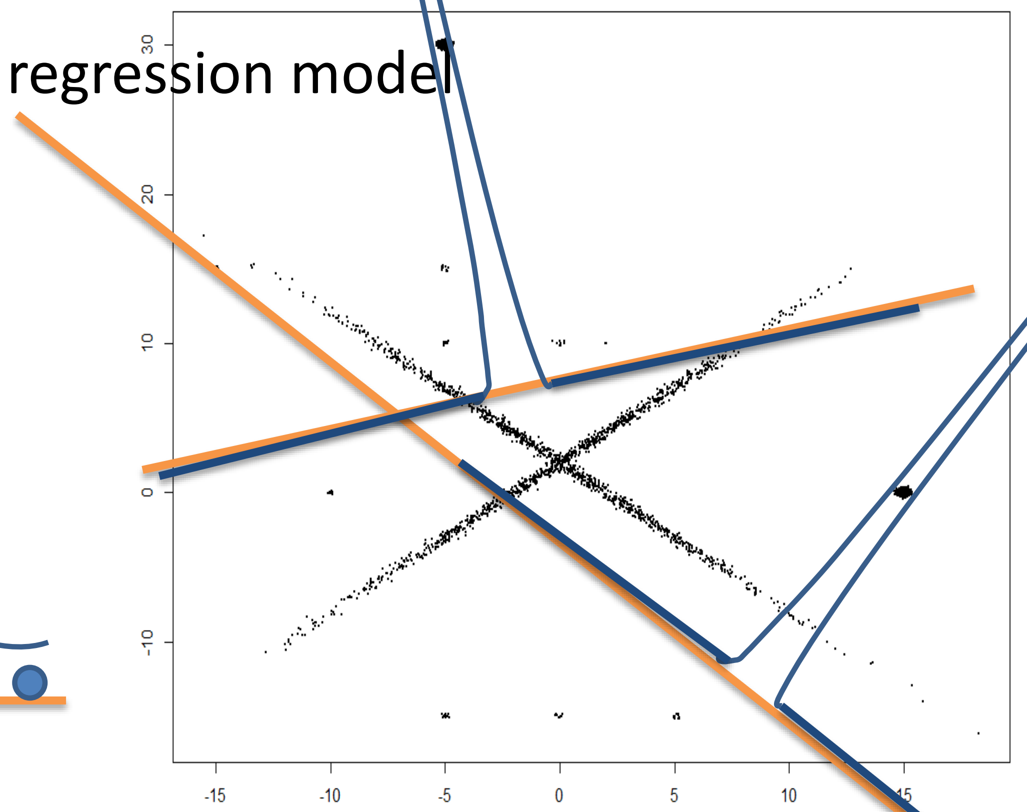
*Assign each observation to the closest model

Estimate density in each model (jointly with all the observations)

Weight based on density

*estimation of each orthogonal regression model





Orthogonal regression - Density based weighting

Start with *k* random linear models

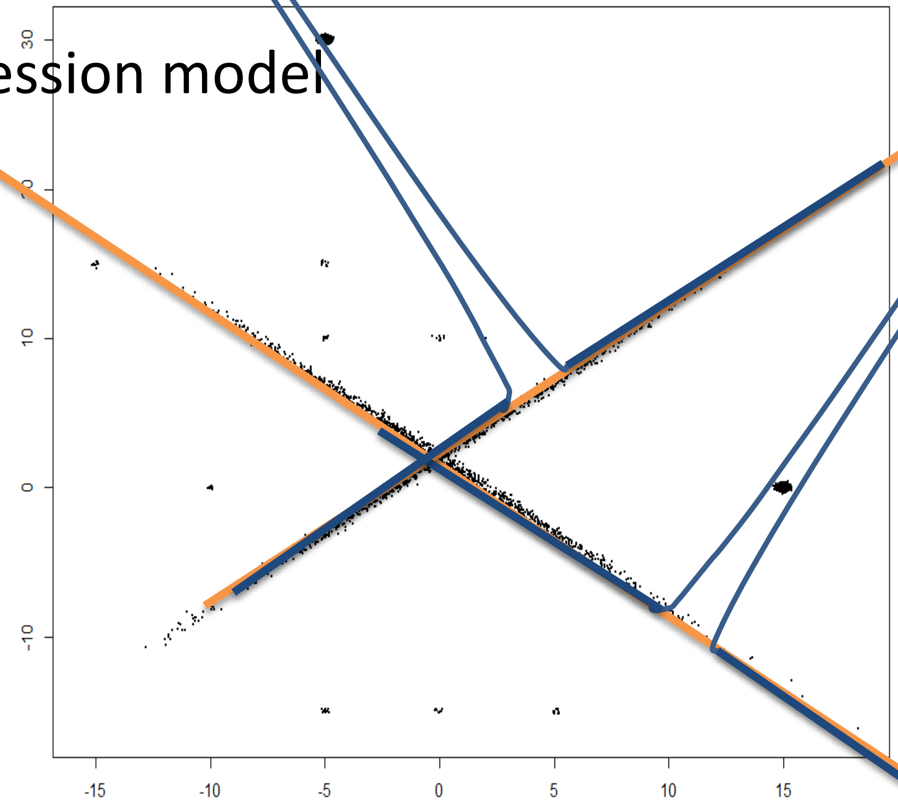
Iterations

*Assign each observation to the closest model

Estimate density in each model

Weight based on density

*estimation of each orthogonal regression model



Pointwise contamination

(40.5%!!) located (0,15)

(40.5%!!) located (-5,30)

(0.09%) in 10 additional locations

Orthogonal regression - Density based weighting

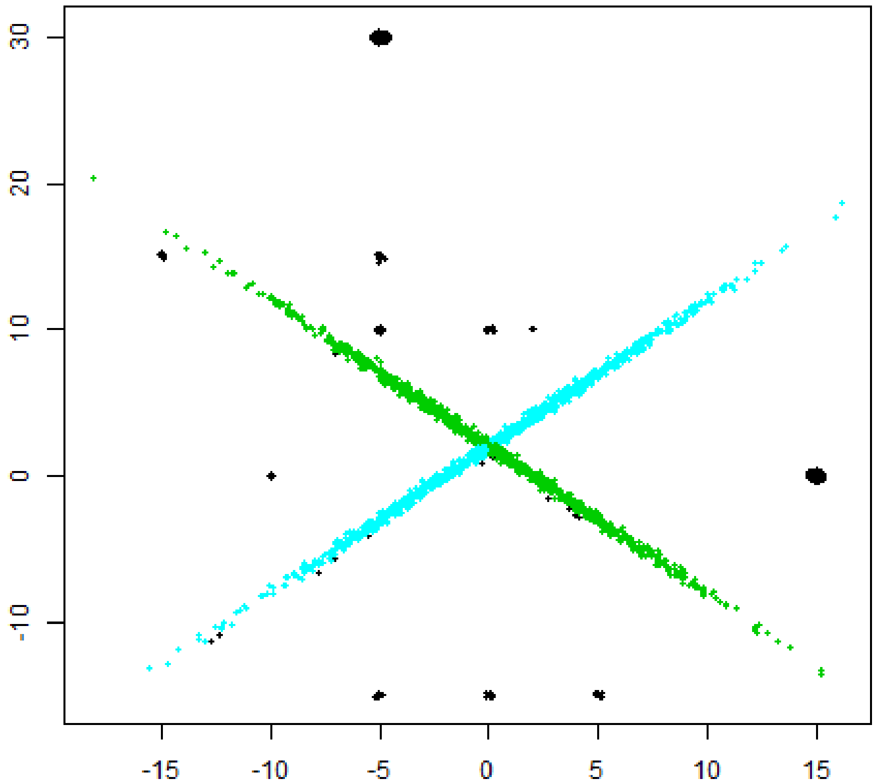
Pointwise contamination

(40.5%!!) located (0,15)

(40.5%!!) located (-5,30)

(0.08%) in 10 additional locations

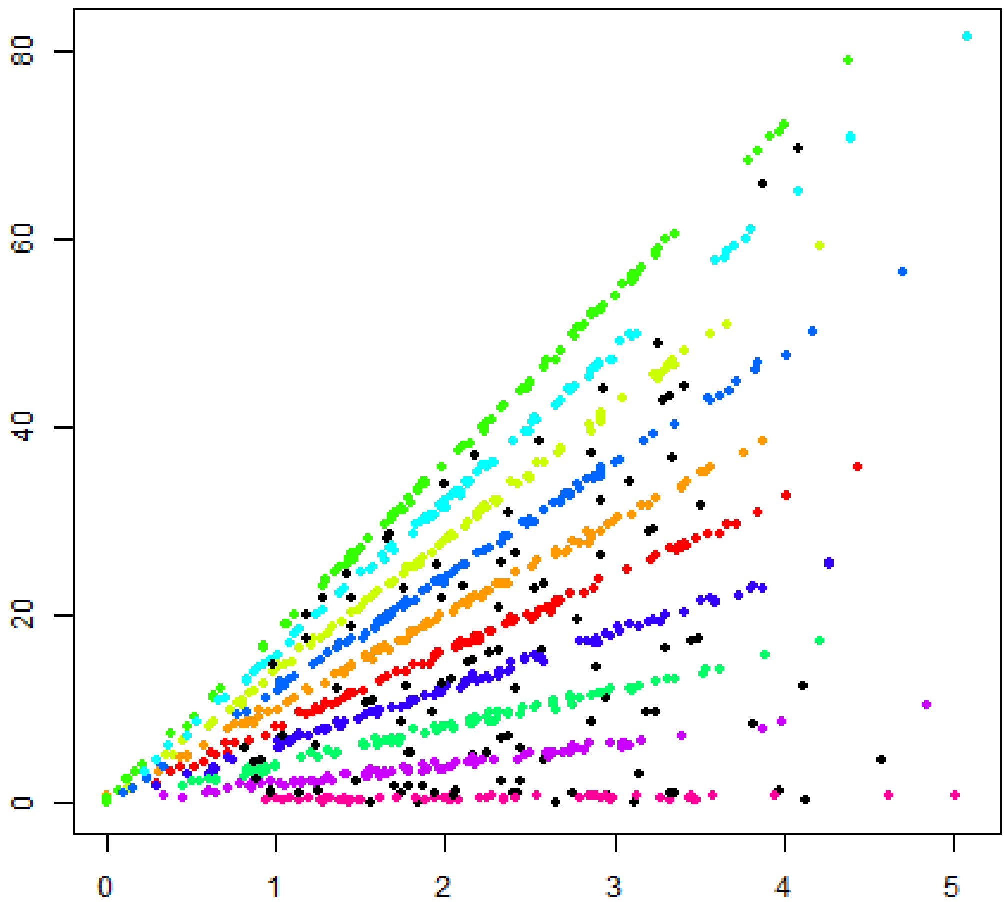
trimming level 30%



Density based weighting

Weighting based on explanatory variables density and estimated beta

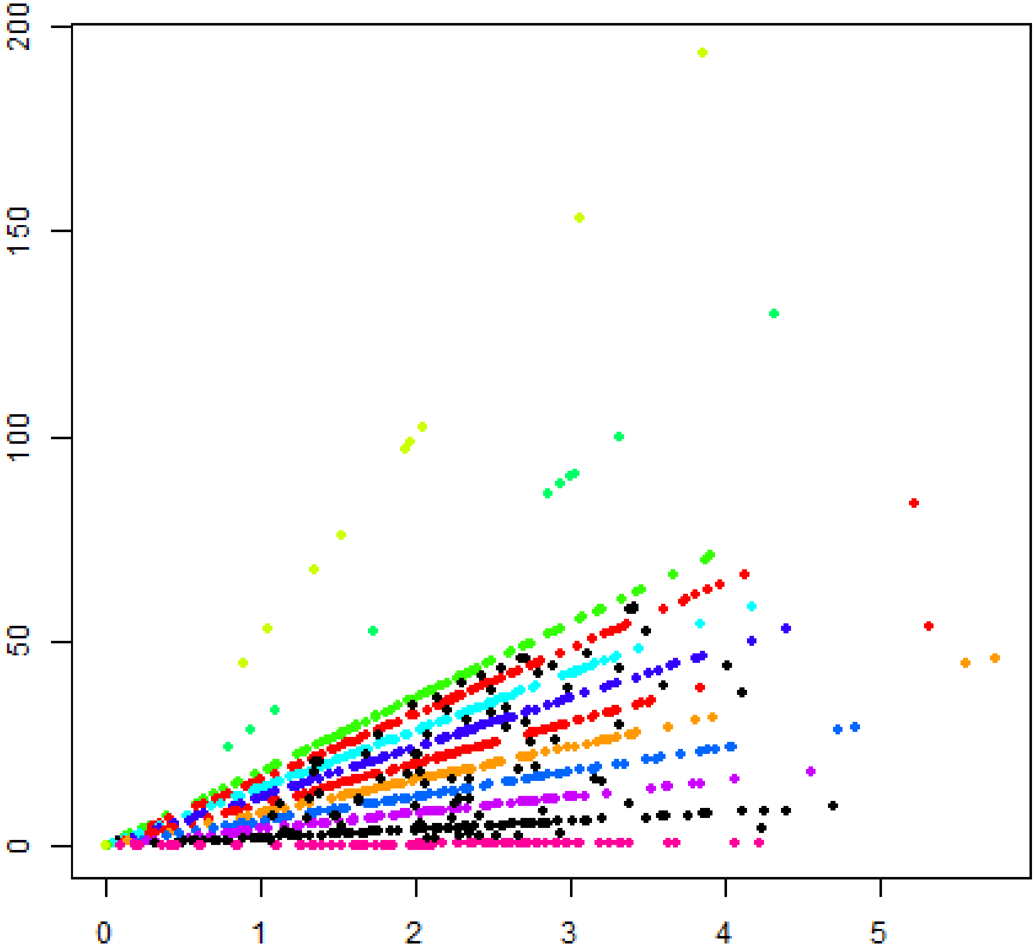
Trimming based on this weighting (not a fix proportion of observations, a fix proporttion of weight)



Density based weighting

Weighting based on explanatory variables density and estimated beta

Trimming based on this weighting (not a fix proportion of observations, a fix proporttion of weight)

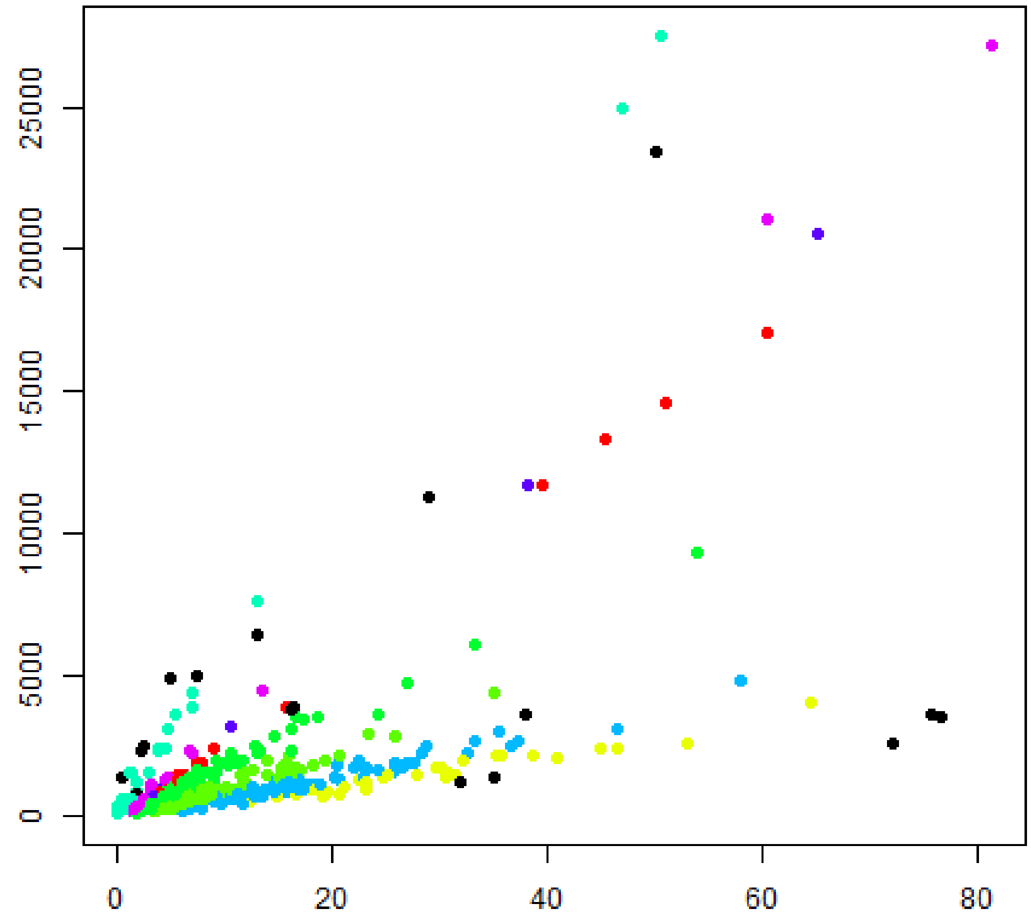


Density based weighting

Weighting based on density and estimated beta

Trimming based on this weighting (not a fix proportion of observations, a fix proporttion of weight)

Product 5810101090



Benford's Law Conference

10-12 July 2019 - Stresa, Italy

Thank you!!!

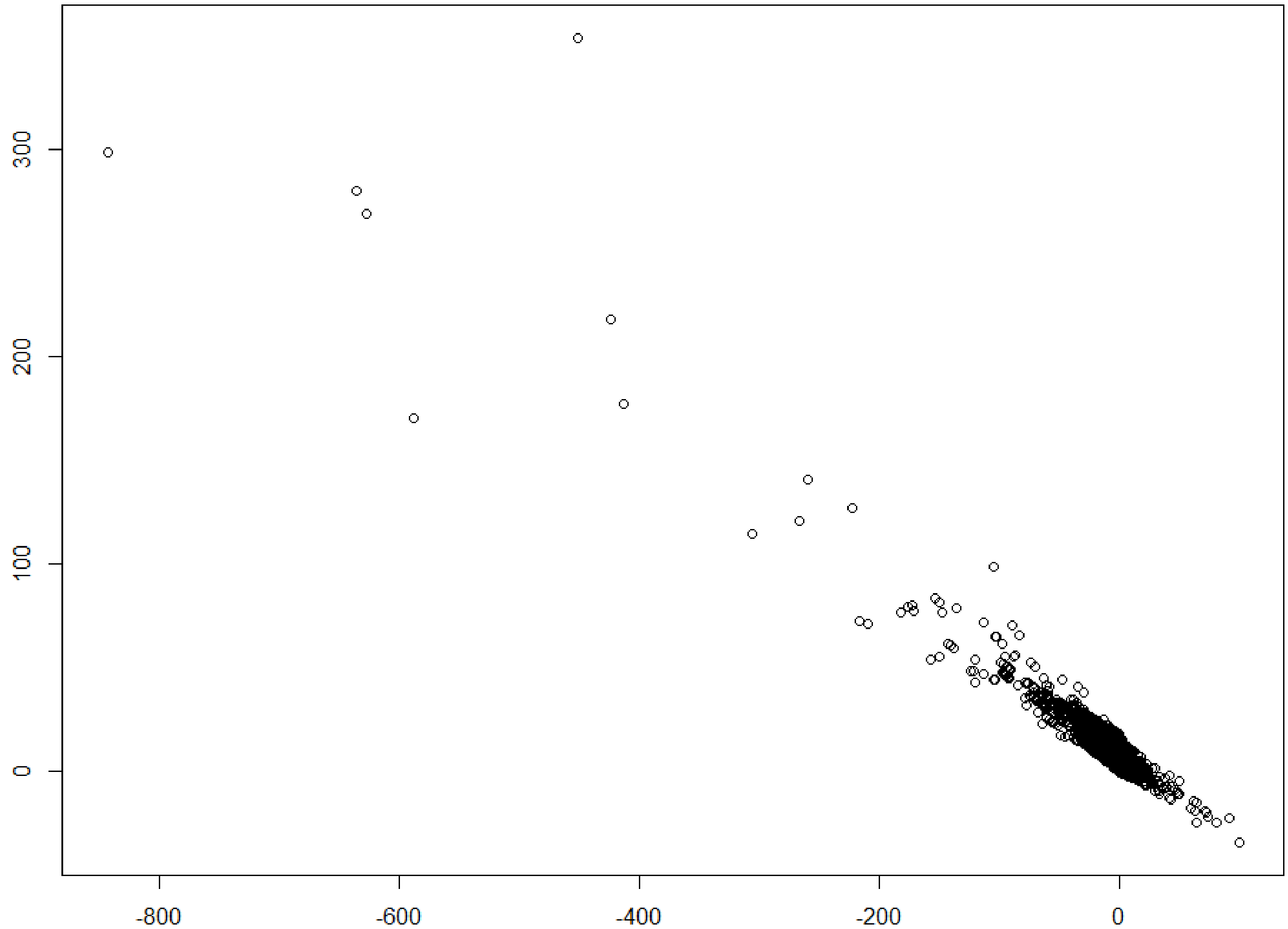
Denoising and Trimming for Improved Cluster Solutions with Applications to Customs Frauds

Andrea Cerioli[1], Luis Ángel García-Escudero[2], Alfonso Gordaliza[2], Carlos Matrán[2], **Agustín Mayo-Iscar[2]**, Domenico Perrotta[3], Marco Riani[1] and Francesca Torti[3]

1. Department of Economics & Ro.S.A. University of Parma
2. Department of Statistics and O.R. & IMUVA. University of Valladolid
3. JRC European Comission. Ispra

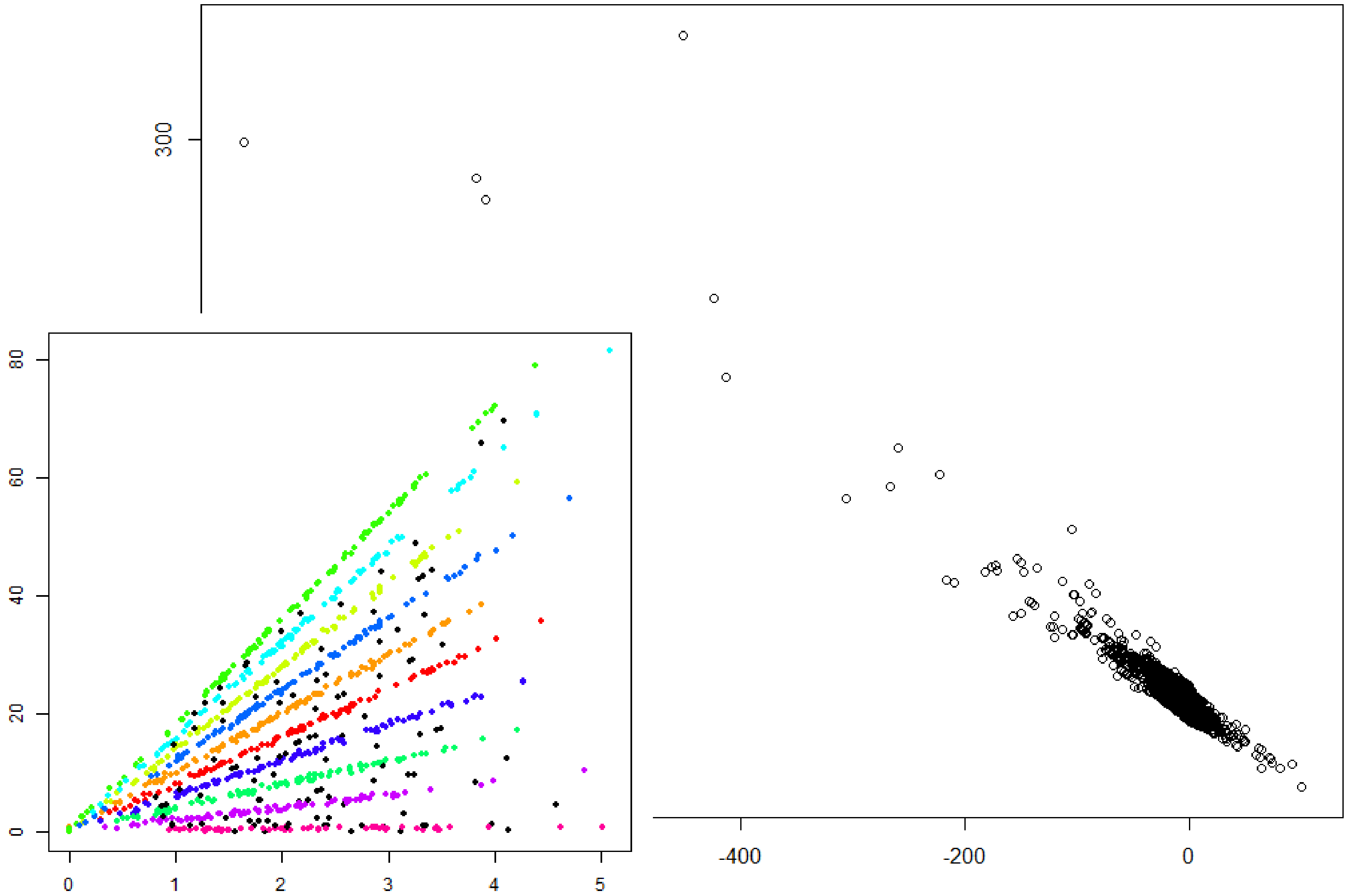
Improving EM results

Estimations for regression parameters obtained in EM runs for different random starts



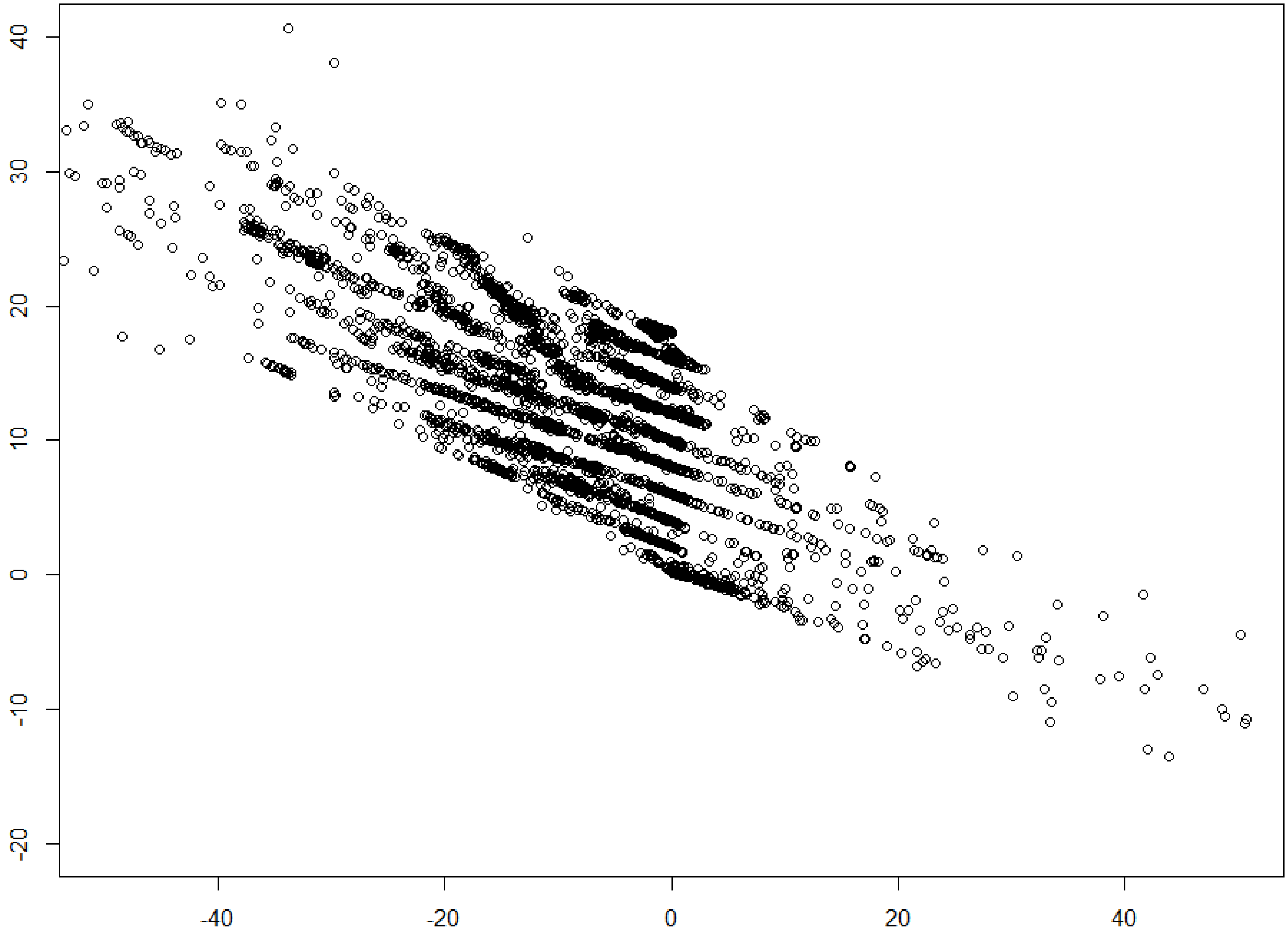
Improving EM results

Estimations for regression parameters obtained in EM runs for different random starts



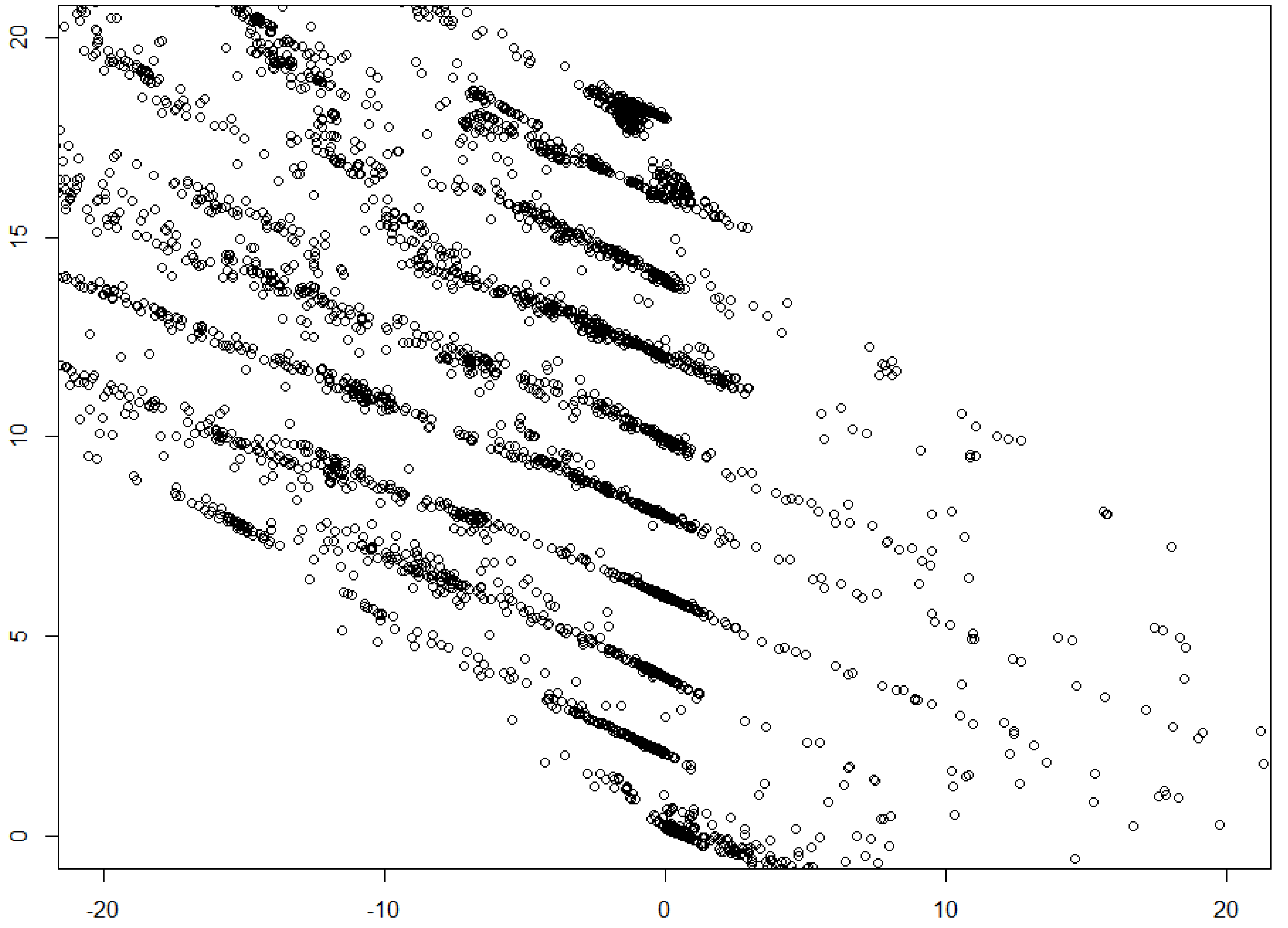
Improving EM results

Estimations for regression parameters obtained in EM runs for different random starts



Improving EM results

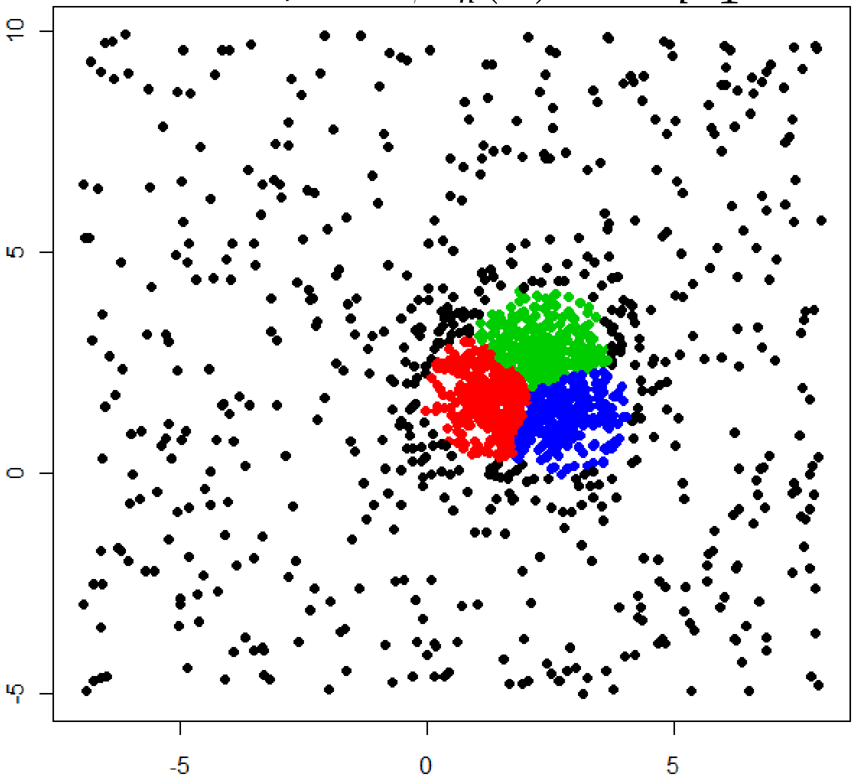
Estimations for regression parameters obtained in EM runs for different random starts



Weights' effect

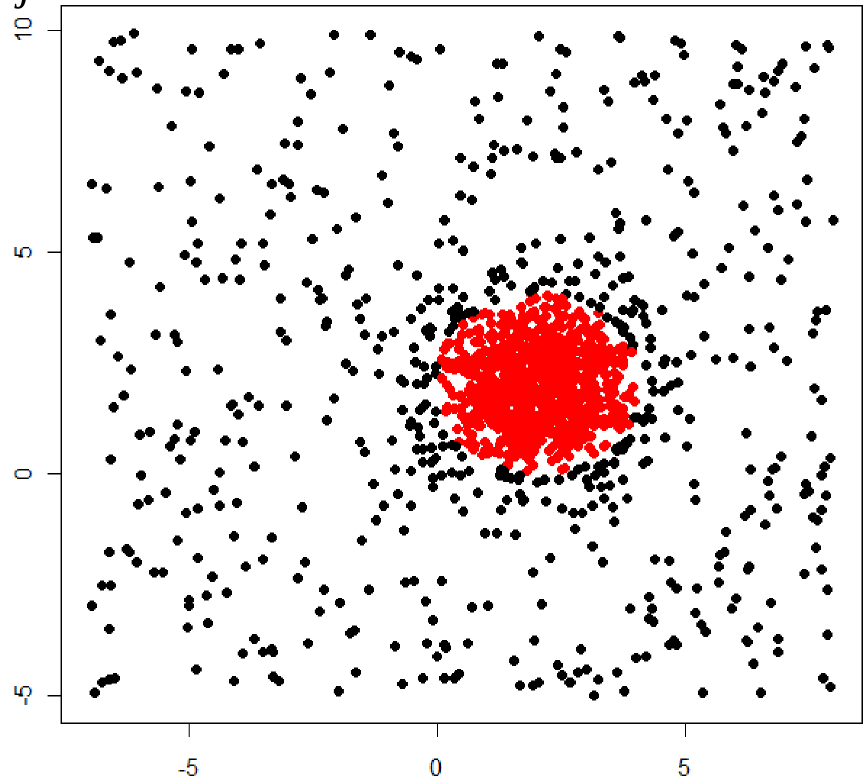
Trimmed k-means $\arg\inf_{\mu}\inf_{z}\inf_{A/P_n(A)=1-\alpha}\sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij} I_A(x_i)\left\|x_i-\mu_j\right\|^2$

$$\arg\sup_{\mu,z}\sup_{A/P_n(A)=1-\alpha}\sum_{i=1}^{n} I_A(x_i)\sum_{j=1}^{k} z_{ij}\log\left(\pi_j N_{\mu_j,\Sigma_j}(x_i)\right)$$



Trimmed k-means *k=3 alpha=20% c=1*

$\pi_1=\ldots=\pi_j=\ldots=\pi_k$



TCLUST *k=3 alpha=20% c=1*

~~$\pi_1=\ldots=\pi_j=\ldots=\pi_k$~~

Input parameters

k: number of clusters
α: level of trimming
c: strength of the constraints

These input parameters have to be provided, in advance, by the user.

Input parameters – k & alpha

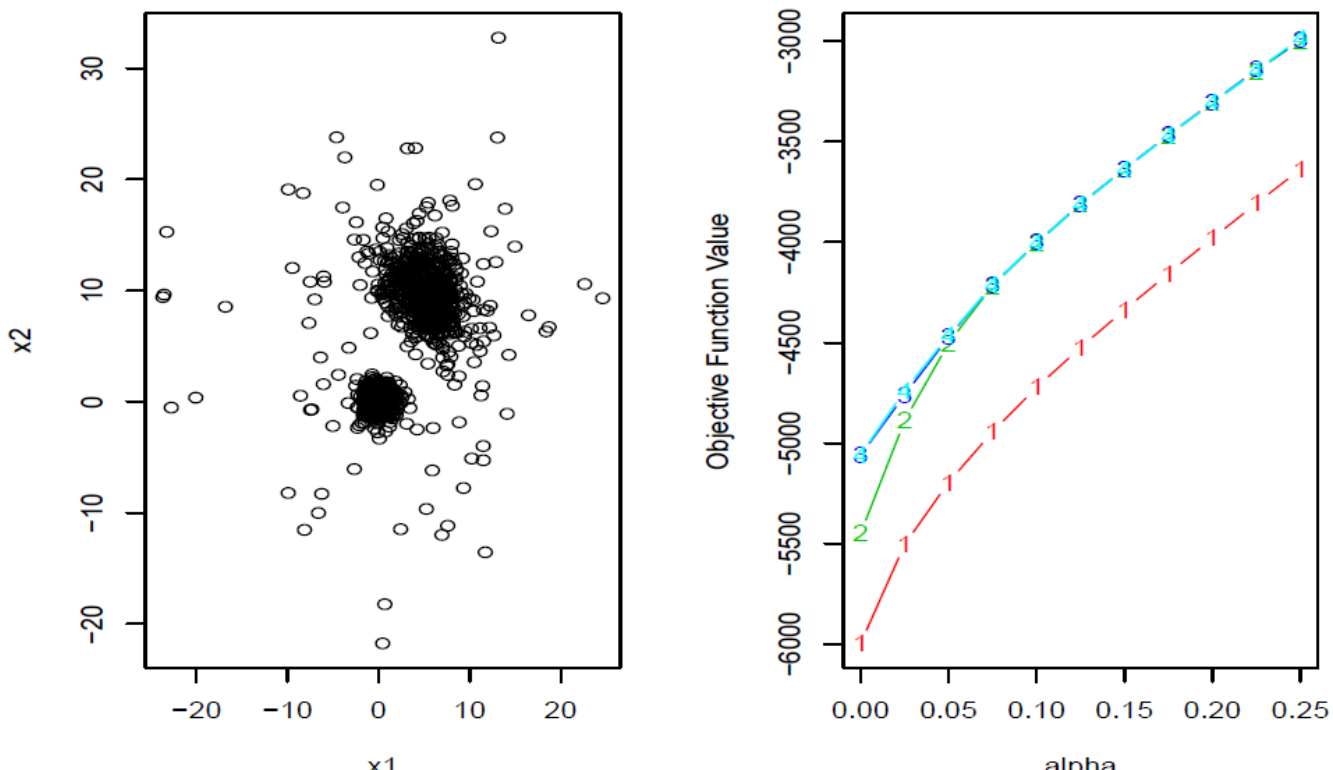
García-Escudero, Gordaliza, Matrán & M-I (2011). Exploring the number of groups in robust model-based clustering. *Stat & Comp*, *21*(4), 585-599.

- **Classification Trimmed Likelihood Curves with** `ctlcurves`**:**

  This tool is given by curves which represents the objective function value for a pairs of (k,alpha).

  These curves can assist users in the selection of values for k and alpha.

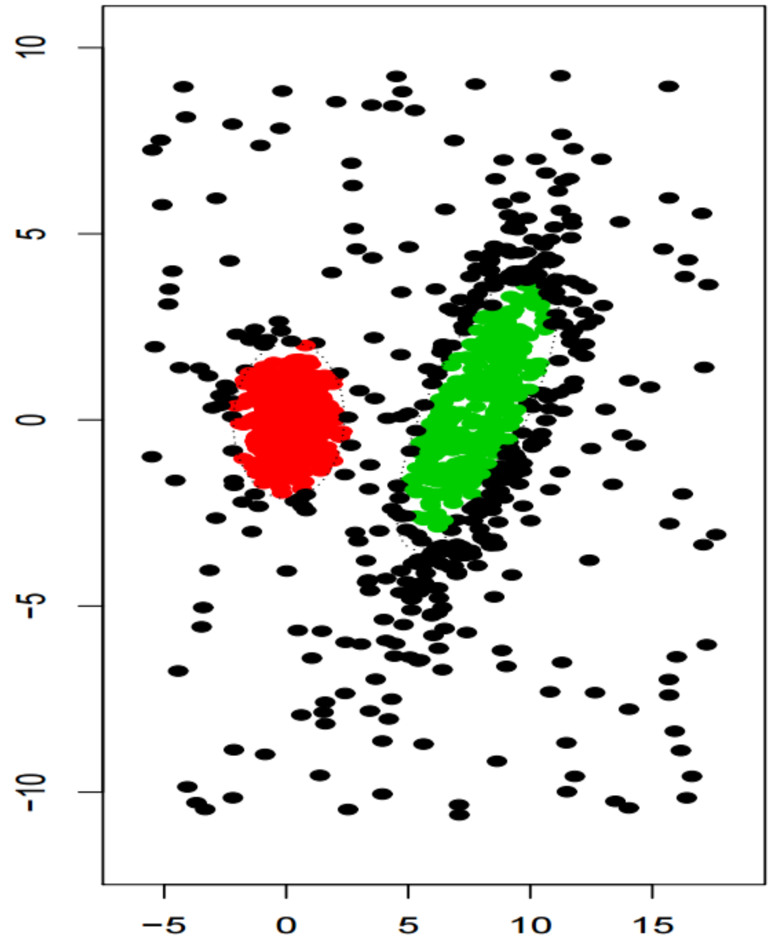
```
> ctlcurves (x,k=1:4,alpha=seq (0,0.25,by = 0.05), restr.fact=100)
```



Input parameters - alpha

ReWeighted TCLUST

Dotto, Farcomeni, García-Escudero & M-I (2018). A Reweighting Approach to Robust Clustering. *Statistics and Computing*.

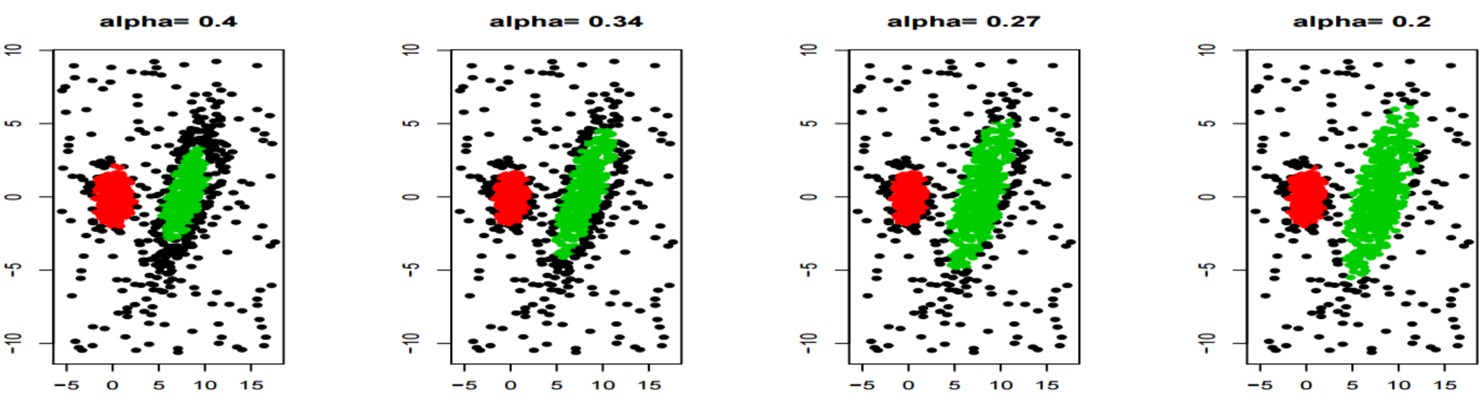


Input parameters - alpha

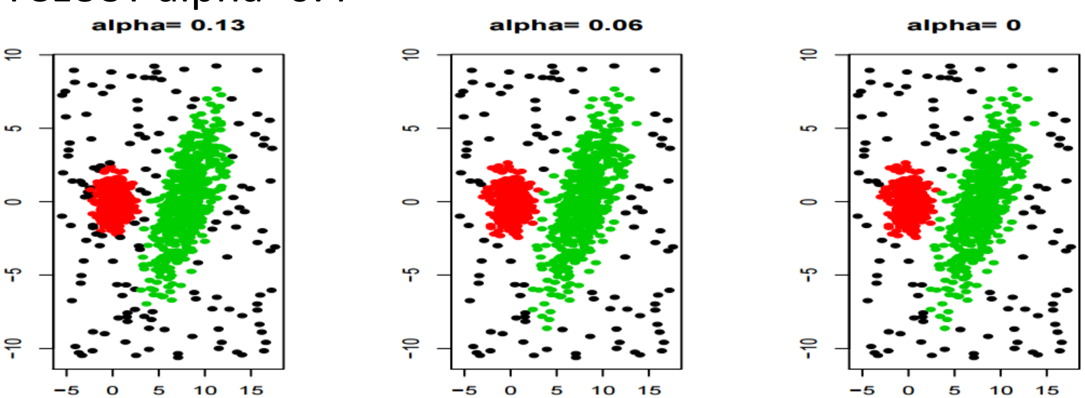
ReWeighted TCLUST

Dotto, Farcomeni, García-Escudero & M-I (2018). A Reweighting Approach to Robust Clustering. *Statistics and Computing*

Starting from a TCLUST solution (high level of trimming), to improve it, sequentially, with reweighting steps



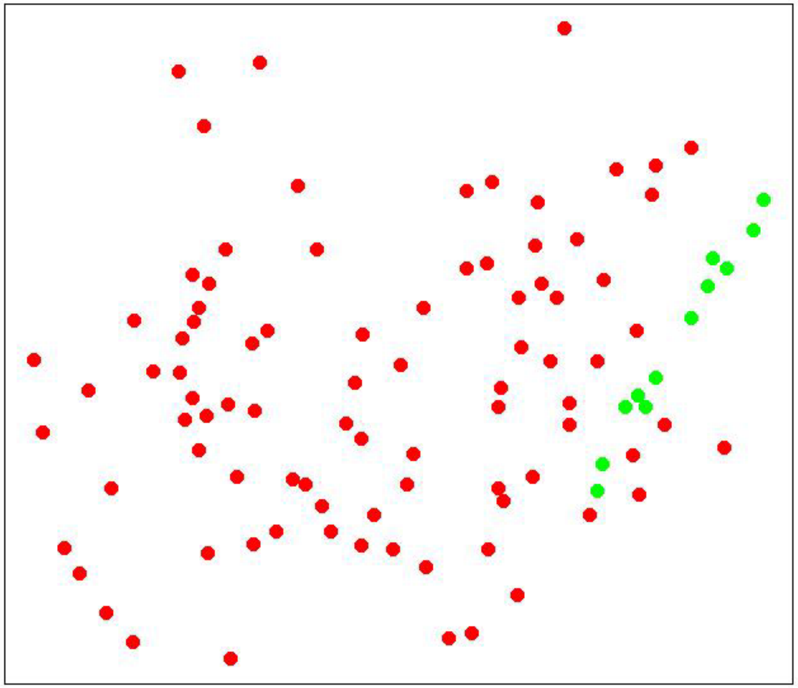
TCLUST alpha=0.4



Input parameters - constraints

Eigenvalue constraints applied to ML finite mixture models estimation

García-Escudero, Gordaliza, Matrán and M-I (2015) Avoiding Spurious Local Maximizers in Mixture Modeling. Stat & Comp, 25 (3) pp 619-633



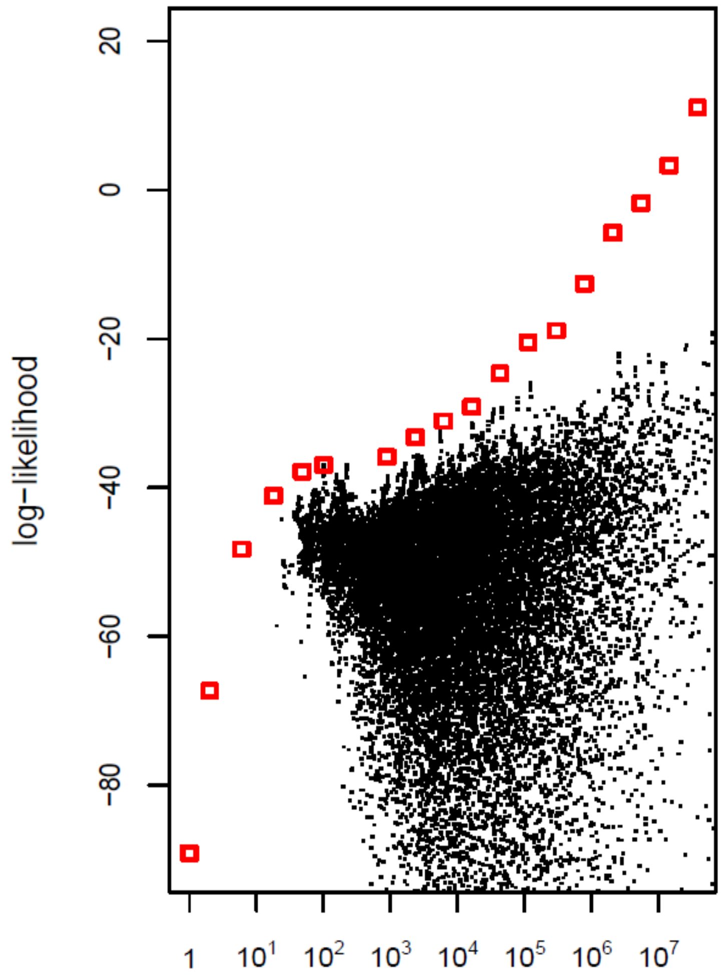
Synthetic data set 2 (McP2000)
Mixture of two normal heteroscedastic populations without contamination

Spurious clusters

- "little practical use or real-world interpretation" (McLachlan &Peel, 2000 McP2000)
- "It often seems in these cases that the model is fitting a small localized random pattern in the data rather than any underlying group structure" . (McP2000)

Constraints for avoiding spurious clusters in the solution

Input parameters - constraints



max of eigenvalue ratios
(log - scale)

Eigenvalue constraints applied to ML finite mixture models estimation

García-Escudero, Gordaliza, Matrán and M-I (2015) Avoiding Spurious Local Maximizers in Mixture Modeling. Stat & Comp, 25 (3) pp 619-633

Prevalence of spurious local maximizers in ML mixture models estimation

When applying the EM algorithm after thousands of random starts, thousands of different solutions appear.

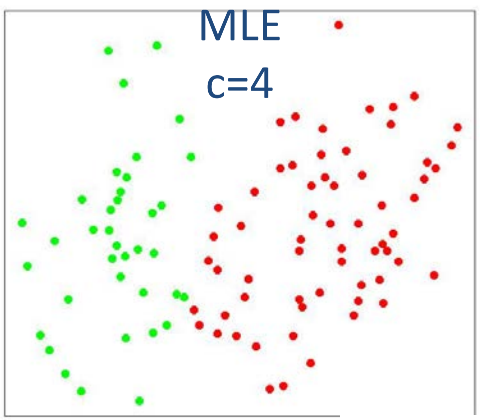
How to choose the estimation?

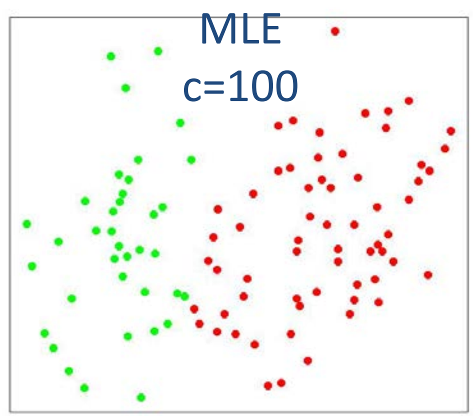
The plot shows Log-likelihoods and maximum of eigenvalues-ratios for thousands of local ML maximizers corresponding to ML estimation of two populations for the virginica species subset of the Iris data.

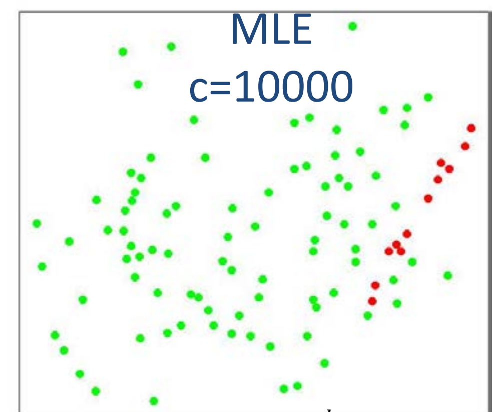
Input parameters - constraints

Eigenvalue contraints applied to ML finite mixture models estimation

García-Escudero, Gordaliza, Matrán and M-I (2015) Avoiding Spurious Local Maximizers in Mixture Modeling. Stat & Comp, 25 (3) pp 619-633

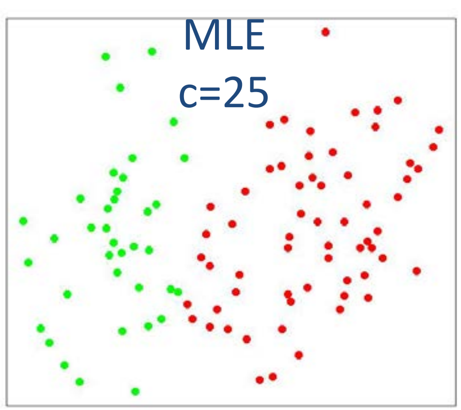


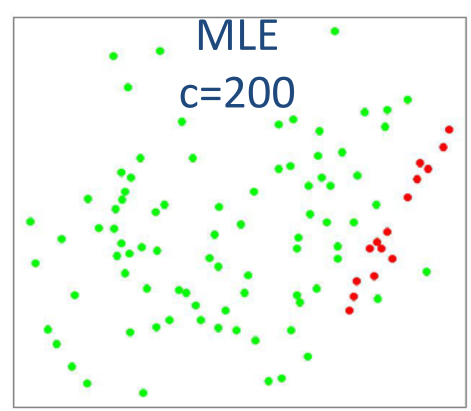




$$\frac{\max \lambda^{l_1}_{\Sigma_{j1}}}{\min \lambda^{l_2}_{\Sigma_{j2}}} \approx 1000$$

$$\frac{\max \lambda^{l_1}_{\Sigma_{j1}}}{\min \lambda^{l_2}_{\Sigma_{j2}}} \approx 3$$





Synthetic data set 2 (McP2000)

Input parameters – k &constraints

BIC based choice of contraints level

Cerioli,A., García-Escudero,L.A., M-I,A., & Riani,M. (2018) Finding the Number of Normal Groups in Model-Based Clustering via Constrained Likelihoods. Journal of Computational and Graphical Statistics, 27(2), 404-416.

Trimmed BIC approach based on monitoring

$(k, c) \to -2\, L(k, \alpha, c) + \nu(k, \alpha, c)$ when α is fixed

where $\nu(k, \alpha, c)$ penalizes a higher (than needed) “model complexities”

Analogous to other BIC approaches but n is replaced by $[n(1 - \alpha)]$

$\nu(k, \alpha, c)$ is an increasing function on c (higher c ⇒ less-constrained Sj matrices ⇒ higher model complexity...)

TCLUST Parsimonious modelling

We are interested in applying trimming & constraints in the estimation of 14 Parsimonious models from Celeux and Govaert (1995). Punzo & McNicholas (2013) gave robust estimators for these models.

We are collaborating with Marco Riani and Andrea Cerioli (University of Parma) in applying trimming & contraints for estimating these models.

Table 1: Nomenclature, covariance structure, and number of free parameters in $\boldsymbol{\Sigma}_g$ for each member of the PMCGD family.

Family	Model	Volume	Shape	Orientation	$\boldsymbol{\Sigma}_g$	# Free parameters in $\boldsymbol{\Sigma}_g$
Spherical	EII	Equal	Spherical	-	$\lambda \boldsymbol{I}$	1
	VII	Variable	Spherical	-	$\lambda_g \boldsymbol{I}$	G
Diagonal	EEI	Equal	Equal	Axis-Align	$\lambda \boldsymbol{\Gamma}$	p
	VEI	Variable	Equal	Axis-Align	$\lambda_g \boldsymbol{\Gamma}$	$G+p-1$
	EVI	Equal	Variable	Axis-Align	$\lambda \boldsymbol{\Gamma}_g$	$1+G(p-1)$
	VVI	Variable	Variable	Axis-Align	$\lambda_g \boldsymbol{\Gamma}_g$	Gp
General	EEE	Equal	Equal	Equal	$\lambda \boldsymbol{\Gamma} \boldsymbol{\Delta} \boldsymbol{\Gamma}'$	$p(p+1)/2$
	VEE	Variable	Equal	Equal	$\lambda_g \boldsymbol{\Gamma} \boldsymbol{\Delta} \boldsymbol{\Gamma}'$	$G+p-1+p(p-1)/2$
	EVE	Equal	Variable	Equal	$\lambda \boldsymbol{\Gamma}_g \boldsymbol{\Delta} \boldsymbol{\Gamma}_g'$	$1+G(p-1)+p(p-1)/2$
	EEV	Equal	Equal	Variable	$\lambda \boldsymbol{\Gamma} \boldsymbol{\Delta}_g \boldsymbol{\Gamma}'$	$p+Gp(p-1)/2$
	VVE	Variable	Variable	Equal	$\lambda_g \boldsymbol{\Gamma}_g \boldsymbol{\Delta} \boldsymbol{\Gamma}_g'$	$Gp+p(p-1)/2$
	VEV	Variable	Equal	Variable	$\lambda_g \boldsymbol{\Gamma} \boldsymbol{\Delta}_g \boldsymbol{\Gamma}'$	$G+p-1+Gp(p-1)/2$
	EVV	Equal	Variable	Variable	$\lambda \boldsymbol{\Gamma}_g \boldsymbol{\Delta}_g \boldsymbol{\Gamma}_g'$	$1+G(p-1)+Gp(p-1)/2$
	VVV	Variable	Variable	Variable	$\lambda_g \boldsymbol{\Gamma}_g \boldsymbol{\Delta}_g \boldsymbol{\Gamma}_g'$	$Gp(p+1)/2$

From Punzo & McNicholas (2013)

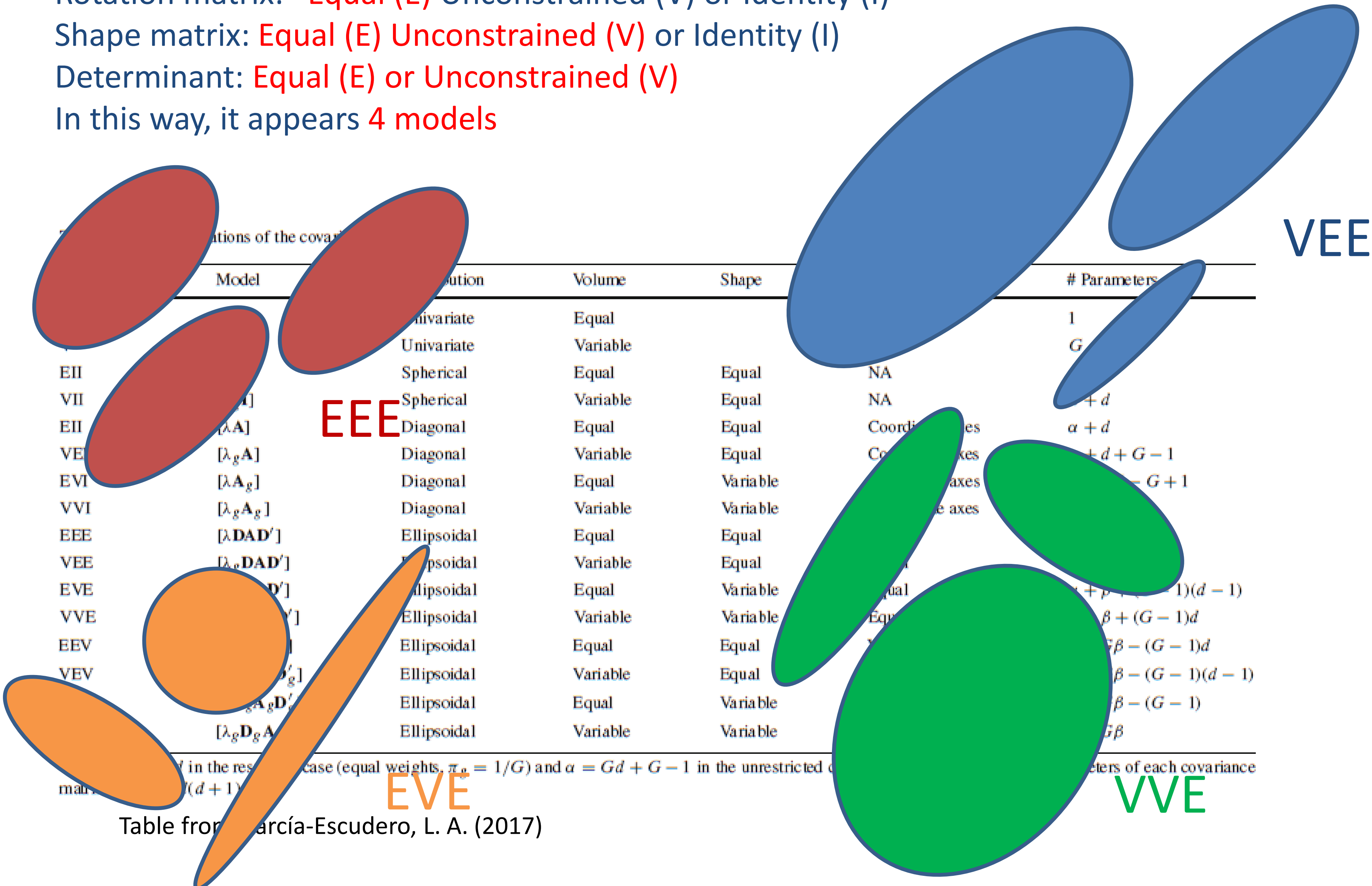
Parsimonious modeling

Rotation matrix: Equal (E) Unconstrained (V) or Identity (I)

Shape matrix: Equal (E) Unconstrained (V) or Identity (I)

Determinant: Equal (E) or Unconstrained (V)

In this way, it appears 4 models



[illegible] ...tions of the cova[illegible]

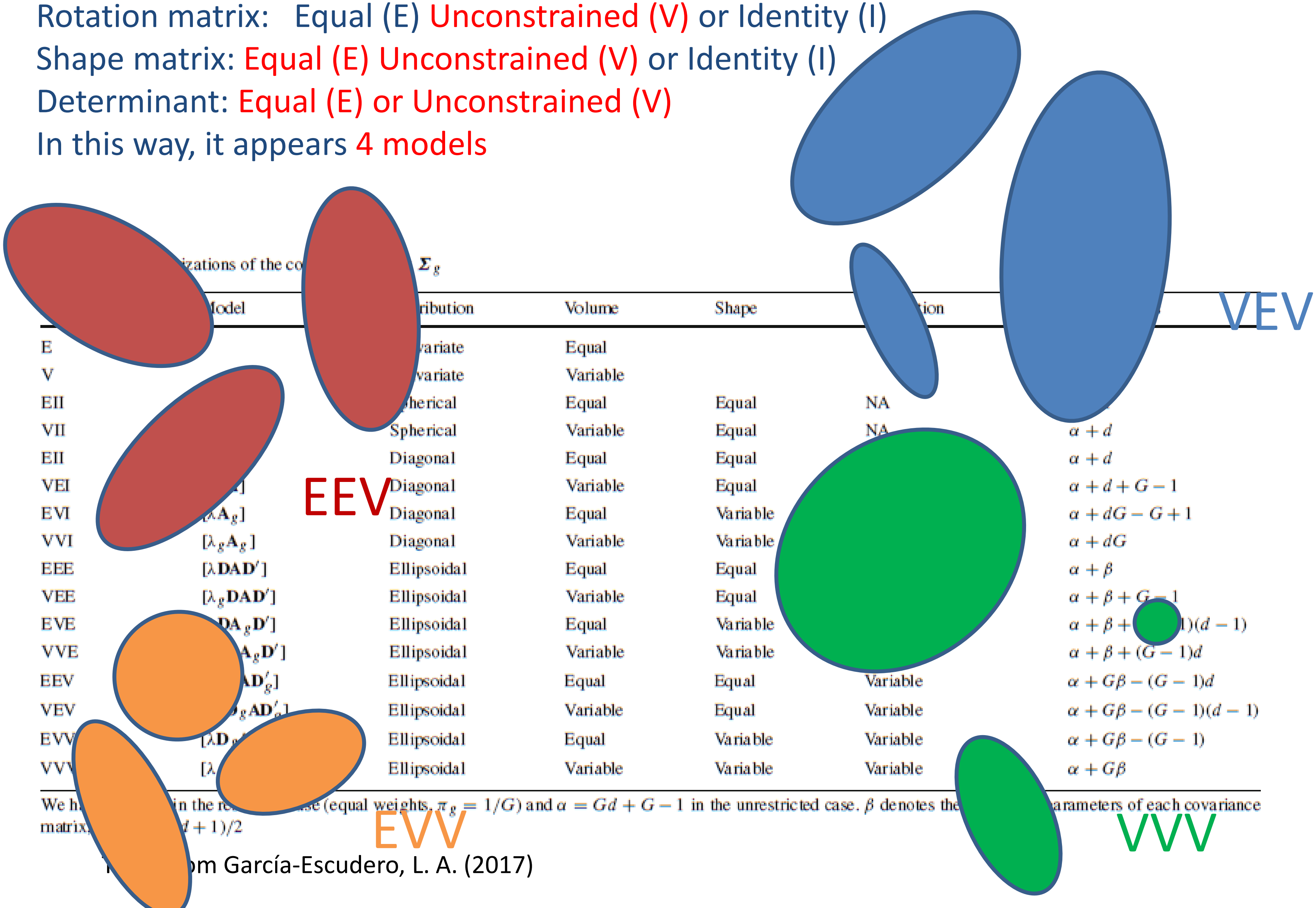
[illegible]	Model	[illegible]ution	Volume	Shape	[illegible]	# Parameters
[illegible]	[illegible]	[illegible]nivariate	Equal			1
[illegible]	[illegible]	Univariate	Variable			G
EII	[illegible]	Spherical	Equal	Equal	NA	[illegible]
VII	[illegible]I]	Spherical	Variable	Equal	NA	[illegible] $+ d$
EII	[λ**A**]	Diagonal	Equal	Equal	Coordi[illegible]es	$\alpha + d$
VE[illegible]	[λ_g**A**]	Diagonal	Variable	Equal	Co[illegible]xes	[illegible] $+ d + G - 1$
EVI	[λ**A**$_g$]	Diagonal	Equal	Variable	[illegible]axes	[illegible] $G + 1$
VVI	[λ_g**A**$_g$]	Diagonal	Variable	Variable	[illegible]e axes	[illegible]
EEE	[λ**DAD**$'$]	Ellipsoidal	Equal	Equal	[illegible]	[illegible]
VEE	[λ_g**DAD**$'$]	[illegible]psoidal	Variable	Equal	[illegible]	[illegible]
EVE	[illegible]**D**$'$]	[illegible]llipsoidal	Equal	Variable	[illegible]ual	[illegible]$1)(d - 1)$
VVE	[illegible]$'$]	Ellipsoidal	Variable	Variable	Equ[illegible]	[illegible]$\beta + (G - 1)d$
EEV	[illegible]	Ellipsoidal	Equal	Equal	[illegible]	[illegible]$G\beta - (G - 1)d$
VEV	[illegible]$_g'$]	Ellipsoidal	Variable	Equal	[illegible]	[illegible]$\beta - (G - 1)(d - 1)$
[illegible]	[illegible]**A**$_g$**D**$'$[illegible]	Ellipsoidal	Equal	Variable	[illegible]	[illegible]$\beta - (G - 1)$
[illegible]	[λ_g**D**$_g$**A**[illegible]	Ellipsoidal	Variable	Variable	[illegible]	[illegible]$G\beta$

[illegible] in the res[illegible] case (equal weights, $\pi_g = 1/G$) and $\alpha = Gd + G - 1$ in the unrestricted [illegible]eters of each covariance ma[illegible] $(d + 1)$[illegible]

Table fro[illegible]arcía-Escudero, L. A. (2017)

Parsimonious modeling

Rotation matrix: Equal (E) Unconstrained (V) or Identity (I)
Shape matrix: Equal (E) Unconstrained (V) or Identity (I)
Determinant: Equal (E) or Unconstrained (V)
In this way, it appears 4 models



...izations of the co... Σ_g

...odel	[illegible]	...ribution	Volume	Shape	...tion	[illegible]
E	[illegible]	...ariate	Equal		[illegible]	[illegible]
V	[illegible]	...variate	Variable		[illegible]	[illegible]
EII	[illegible]	...herical	Equal	Equal	NA	[illegible]
VII	[illegible]	Spherical	Variable	Equal	NA	$\alpha + d$
EII	[illegible]	Diagonal	Equal	Equal	[illegible]	$\alpha + d$
VEI	...]	Diagonal	Variable	Equal	[illegible]	$\alpha + d + G - 1$
EVI	$[\lambda \mathbf{A}_g]$	Diagonal	Equal	Variable	[illegible]	$\alpha + dG - G + 1$
VVI	$[\lambda_g \mathbf{A}_g]$	Diagonal	Variable	Variable	[illegible]	$\alpha + dG$
EEE	$[\lambda \mathbf{DAD}']$	Ellipsoidal	Equal	Equal	[illegible]	$\alpha + \beta$
VEE	$[\lambda_g \mathbf{DAD}']$	Ellipsoidal	Variable	Equal	[illegible]	$\alpha + \beta + G$ ...1
EVE	...$\mathbf{DA}_g\mathbf{D}']$	Ellipsoidal	Equal	Variable	[illegible]	$\alpha + \beta +$ ...$1)(d - 1)$
VVE	...$\mathbf{A}_g\mathbf{D}']$	Ellipsoidal	Variable	Variable	[illegible]	$\alpha + \beta + (G - 1)d$
EEV	...$\mathbf{AD}_g']$	Ellipsoidal	Equal	Equal	Variable	$\alpha + G\beta - (G - 1)d$
VEV	...$_g\mathbf{AD}_g'$...]	Ellipsoidal	Variable	Equal	Variable	$\alpha + G\beta - (G - 1)(d - 1)$
EVV	$[\lambda \mathbf{D}_g$...	Ellipsoidal	Equal	Variable	Variable	$\alpha + G\beta - (G - 1)$
VVV	$[\lambda$...	Ellipsoidal	Variable	Variable	Variable	$\alpha + G\beta$

We h... in the re... se (equal weights, $\pi_g = 1/G$) and $\alpha = Gd + G - 1$ in the unrestricted case. β denotes the ... arameters of each covariance matrix, ...$d + 1)/2$

T... om García-Escudero, L. A. (2017)

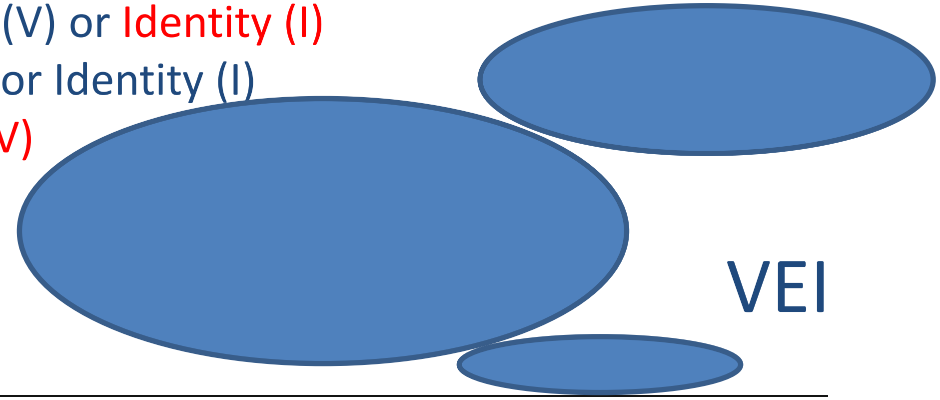
Parsimonious modeling

Rotation matrix: Equal (E) Unconstrained (V) or Identity (I)

Shape matrix: Equal (E) Unconstrained (V) or Identity (I)

Determinant: Equal (E) or Unconstrained (V)

In this way, it appears 4 models



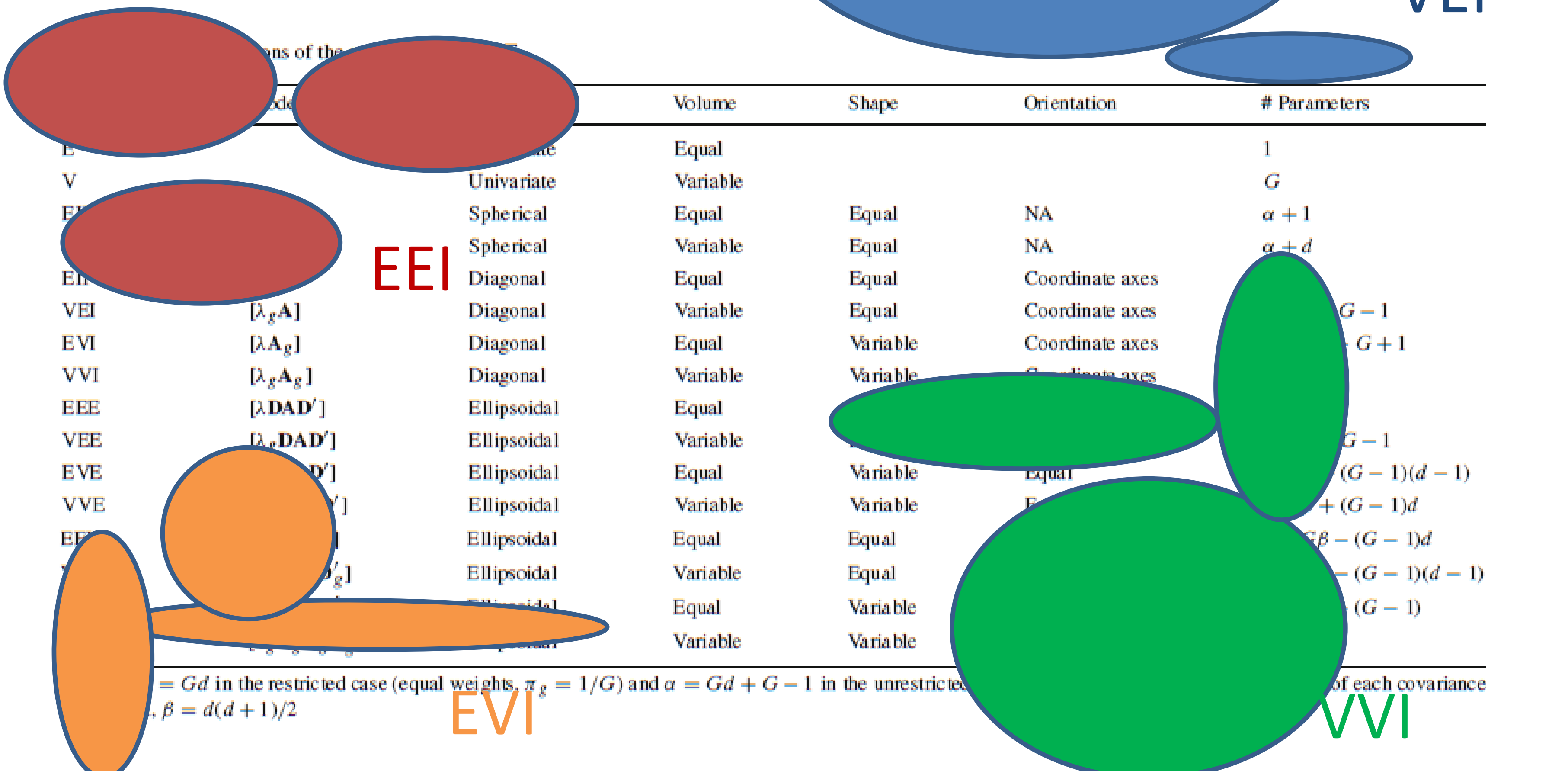


Table from García-Escudero, L. A. (2017)

Parsimonious modeling

Rotation matrix: Equal (E) Unconstrained (V) or Identity (I)
Shape matrix: Equal (E) Unconstrained (V) or Identity (I)
Determinant: Equal (E) or Unconstrained (V)
In this way, it appears 2 models

izations of the covarianc

[illegible]	Model	[illegible]	Volume	Shape	Orientation	# Parameters
[illegible]	[illegible]	[illegible]	Equal			1
[illegible]	[illegible]	[illegible]	Variable			[illegible]
EII	[illegible]	[illegible]	Equal	Equal	NA	[illegible]
VII	[illegible]	Spherica	Variable	Equ	NA	[illegible]
EII	[illegible]	Diagonal	Equal	[illegible]	Coordinate axes	[illegible]
VEI	[illegible]	Diagonal	Variable	[illegible]	Coordinate axes	$\alpha + d + G - 1$
EVI	[illegible]	Diagonal	Equal	[illegible]	oordinate axes	$\alpha + dG - G + 1$
VVI	[illegible]	Diagonal	Variable	[illegible]	oordinate axes	$\alpha + dG$
EEE	$\mathbf{D}']$	Ellipsoidal	Equal	[illegible]	Equal	$\alpha + \beta$
VEE	$[\lambda_g \mathbf{DAD}']$	Ellipsoidal	Variable	[illegible]	Equal	$\alpha + \beta + G - 1$
EVE	$[\lambda \mathbf{DA}_g\mathbf{D}']$	Ellipsoidal	Equal	Varia	Equal	$+ \beta + (G-1)(d-1)$
VVE	$[\lambda_g \mathbf{DA}_g\mathbf{D}']$	Ellipsoidal	Variable	Variable	Equal	$+ \beta + (G-1)d$
EEV	$[\lambda \mathbf{D}_g\mathbf{AD}_g']$	Ellipsoidal	Equal	Equal	Variable	$+ G\beta - (G-1)d$
VEV	$[\lambda_g \mathbf{D}_g\mathbf{AD}_g']$	Ellipsoidal	Variable	Equal	Variable	$\alpha + G\beta - (G-1)(d-1)$
EVV	$[\lambda \mathbf{D}_g\mathbf{A}_g\mathbf{D}_g']$	Ellipsoidal	Equal	Variable	Variable	$\alpha + G\beta - (G-1)$
VVV	$[\lambda_g \mathbf{D}_g\mathbf{A}_g\mathbf{D}_g']$	Ellipsoidal	Variable	Variable	Variable	$\alpha + G\beta$

We have $\alpha = Gd$ in the restricted case (equal weights, $\pi_g = 1/G$) and $\alpha = Gd + G - 1$ in the unrestricted case. β denotes the number of parameters of each covariance matrix, i.e., $\beta = d(d+1)/2$

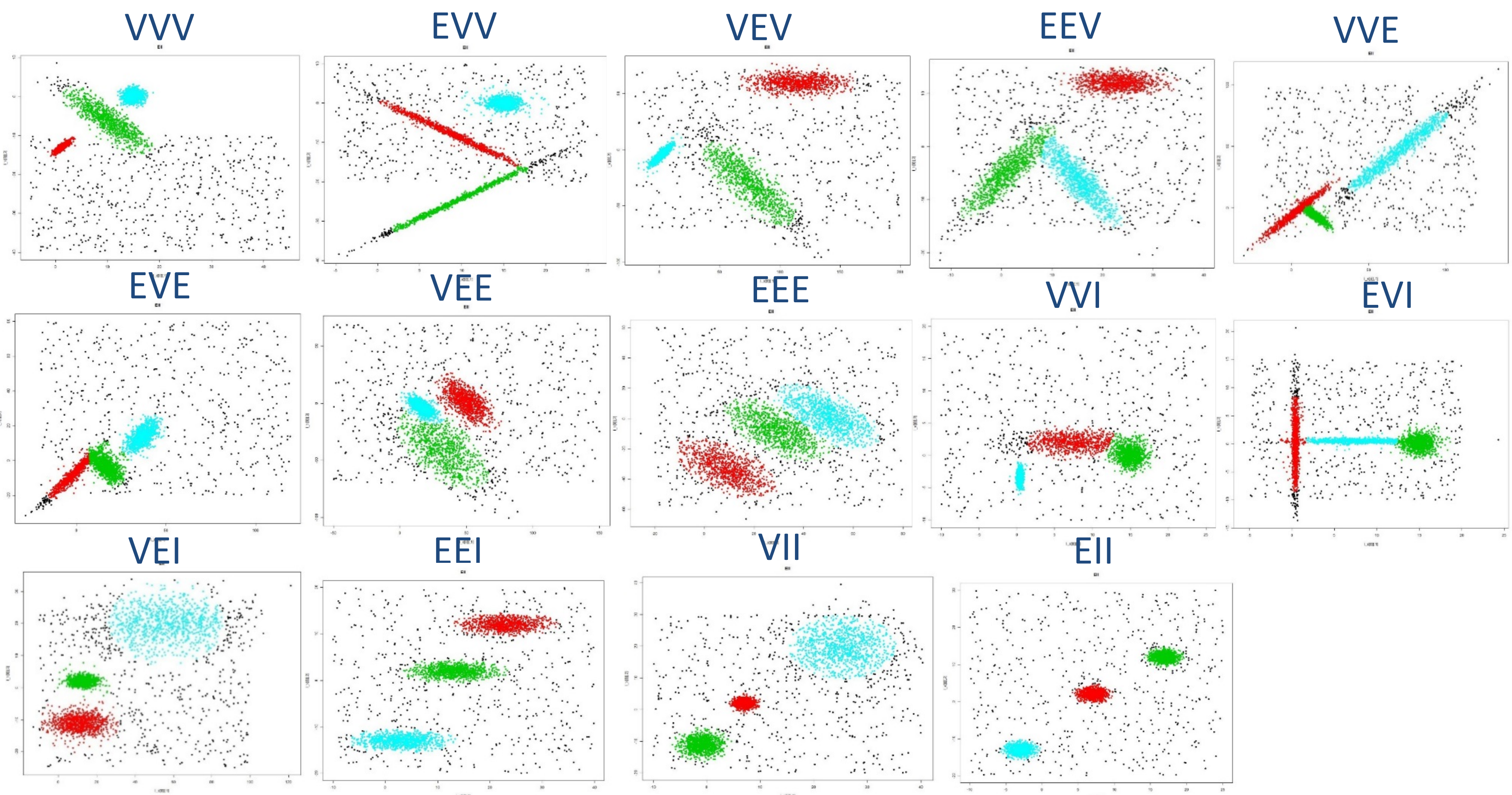
EII

VII

Table from García-Escudero, L. A. (2017)

TCLUST Parsimonious modelling

Trimming & constraints for estimating 14 Parsimonious models from Celeux and Govaert (1995)



Trimmed & constrained Estimation applied to artificial data from the corresponding model + contamination $\alpha = 15\%$ $c=25$

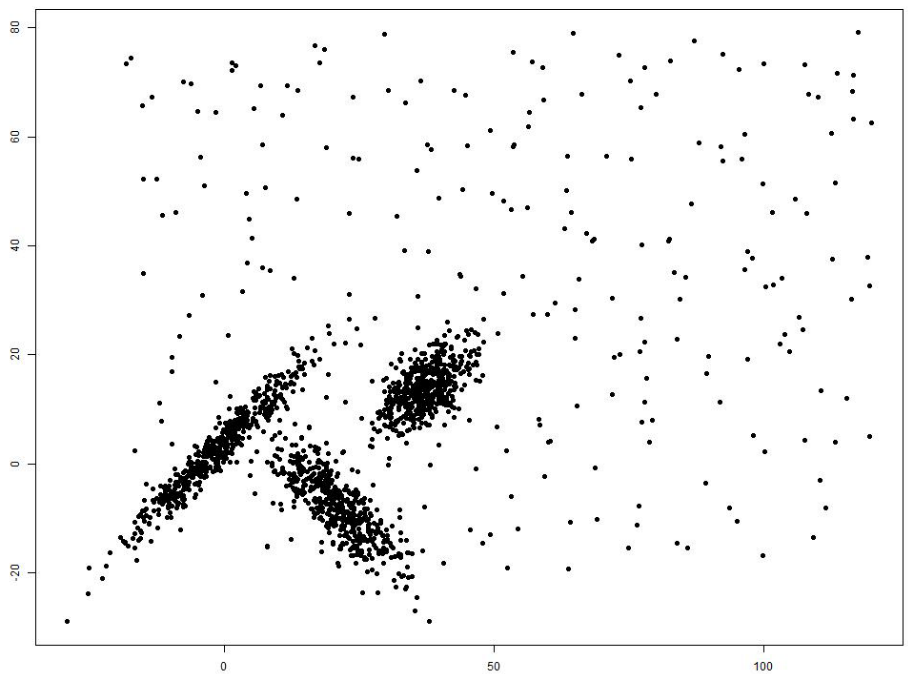
TCLUST Parsimonious modelling

Trimming & constraints for estimating 14 Parsimonious models from Celeux and Govaert (1995)

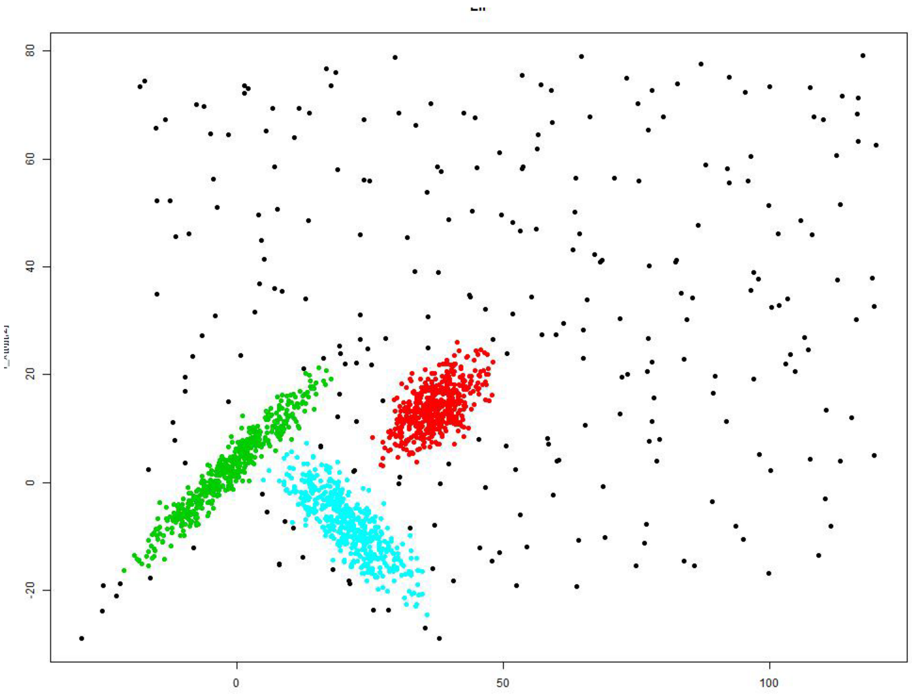
BIC penalized TCLUST estimation α= 15% applied to artificial data from VVE model

Search in 14 models for k=1,2,3,4,5

from VVE K=3



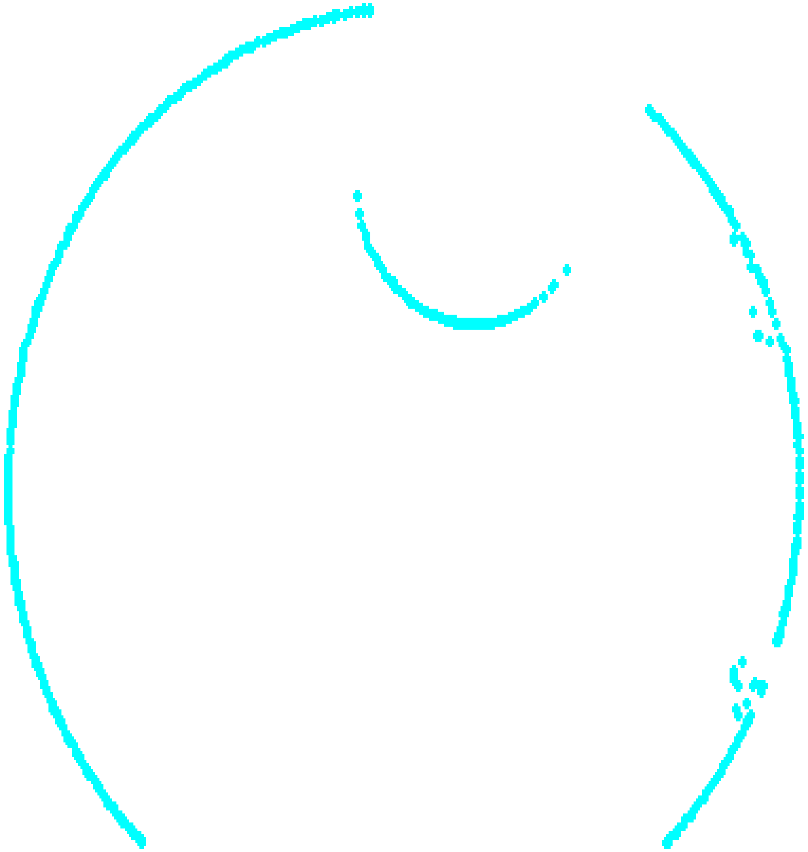
estimation: VVE k=3



Clustering around parametrical curves

García-Escudero, M-I, Sánchez-Gutiérrez (CSDA, 2017)

Train tunnel



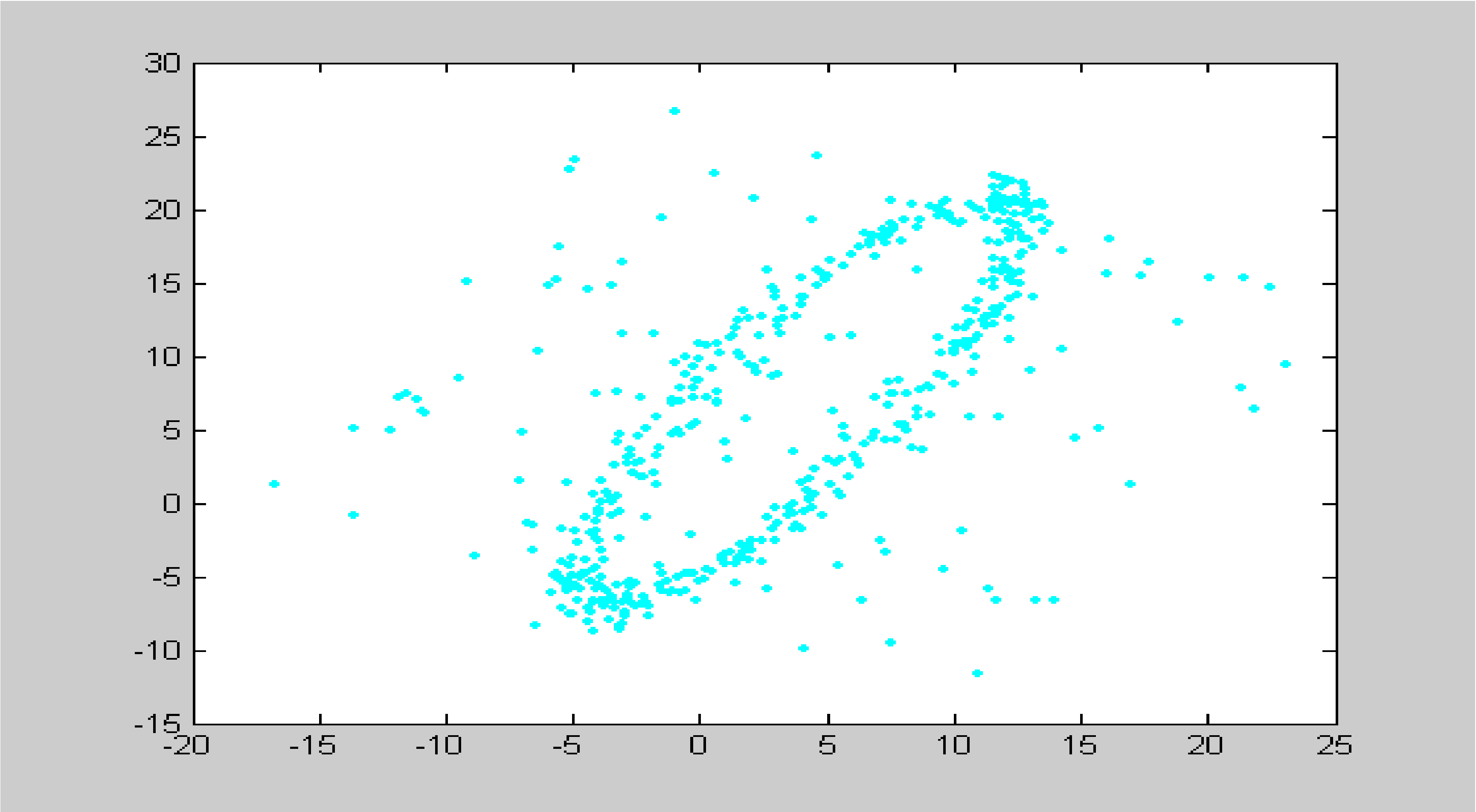
Retinography

Clustering around parametrical curves

García-Escudero, M-I, Sánchez-Gutiérrez (CSDA, 2017)

ellipse

contamination=0.20

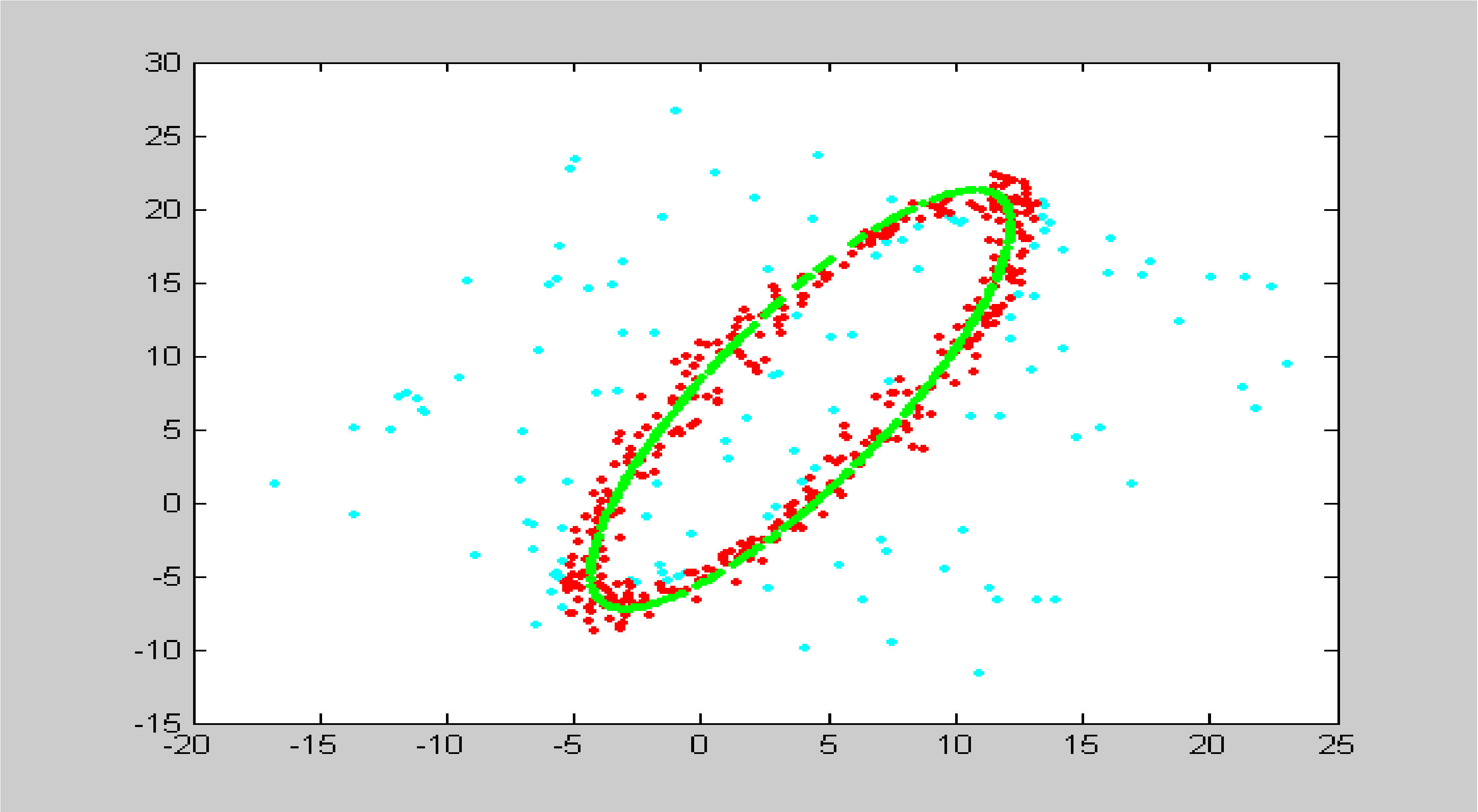


Clustering around parametrical curves

García-Escudero, M-I, Sánchez-Gutiérrez (CSDA, 2017)

ellipse

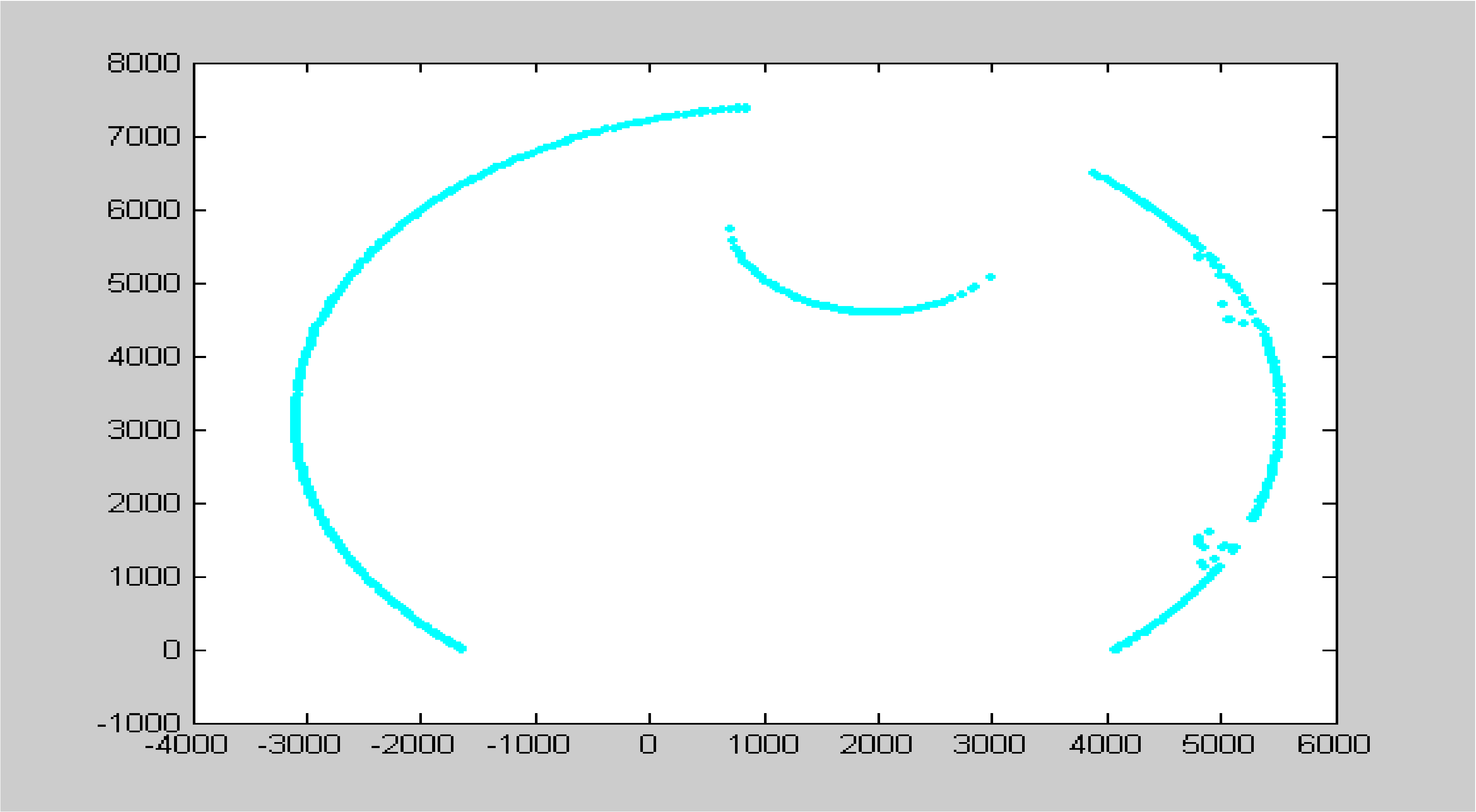
contamination=0.20 trimming level =0.25



Clustering around parametrical curves

García-Escudero, M-I, Sánchez-Gutiérrez (CSDA, 2017)

High speed train tunnel (ellipse)

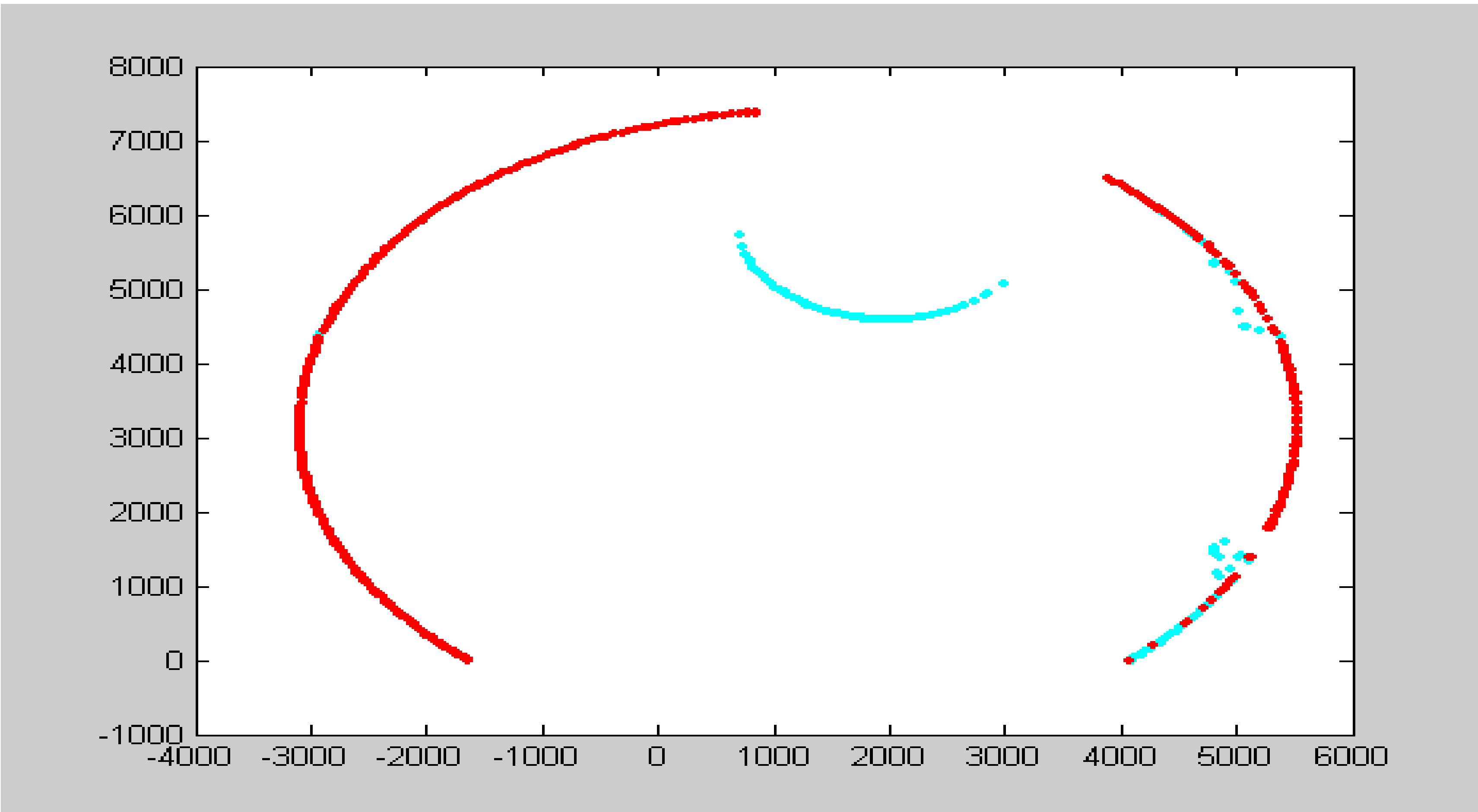


Clustering around parametrical curves

García-Escudero, M-I, Sánchez-Gutiérrez (CSDA, 2017)

High speed train tunnel (ellipse)

trimming level =0.25



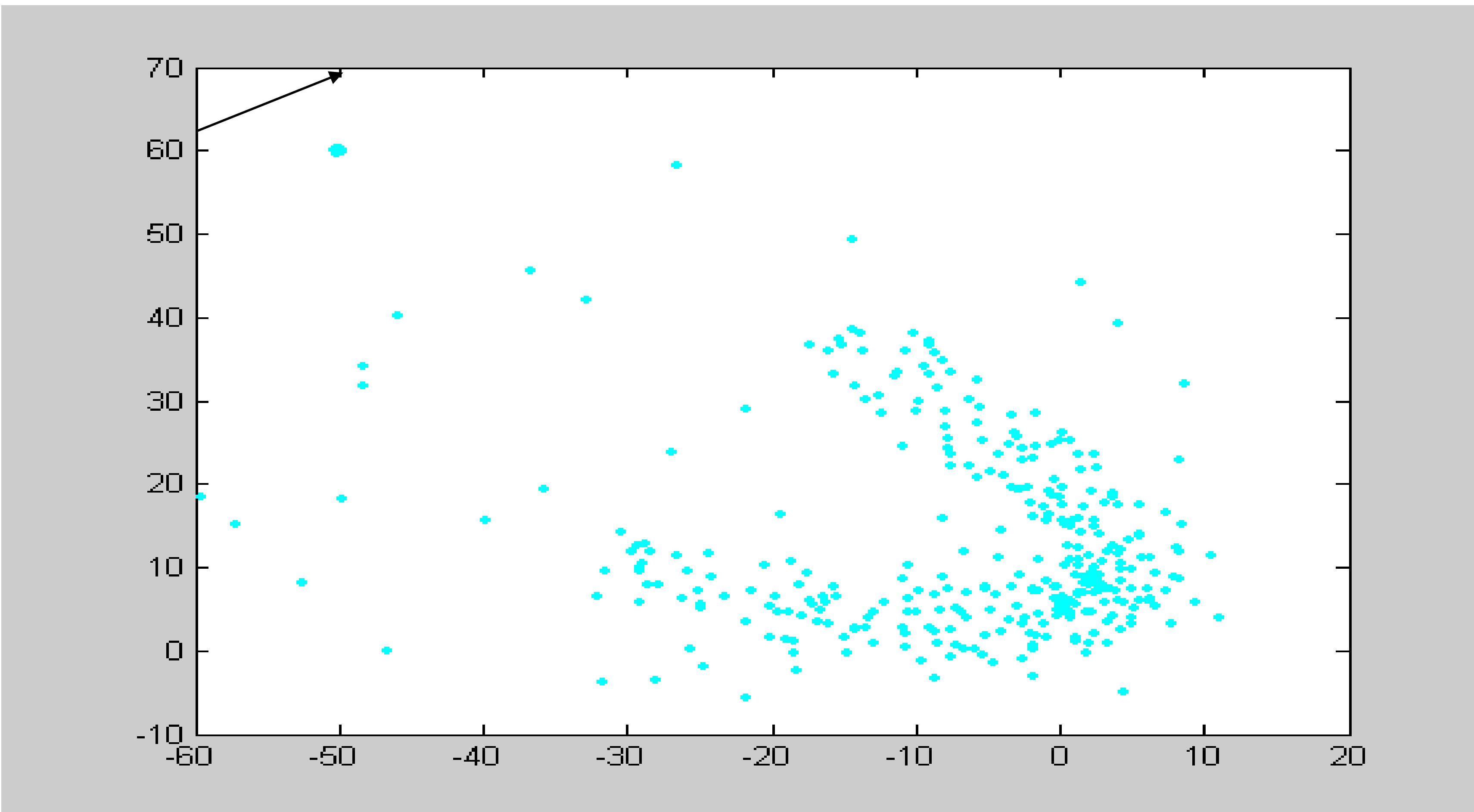
trimming level =0.25

Clustering around parametrical curves

García-Escudero, M-I, Sánchez-Gutiérrez (CSDA, 2017)

Parabola

background noise=0.08 pointwise contamination=0.16

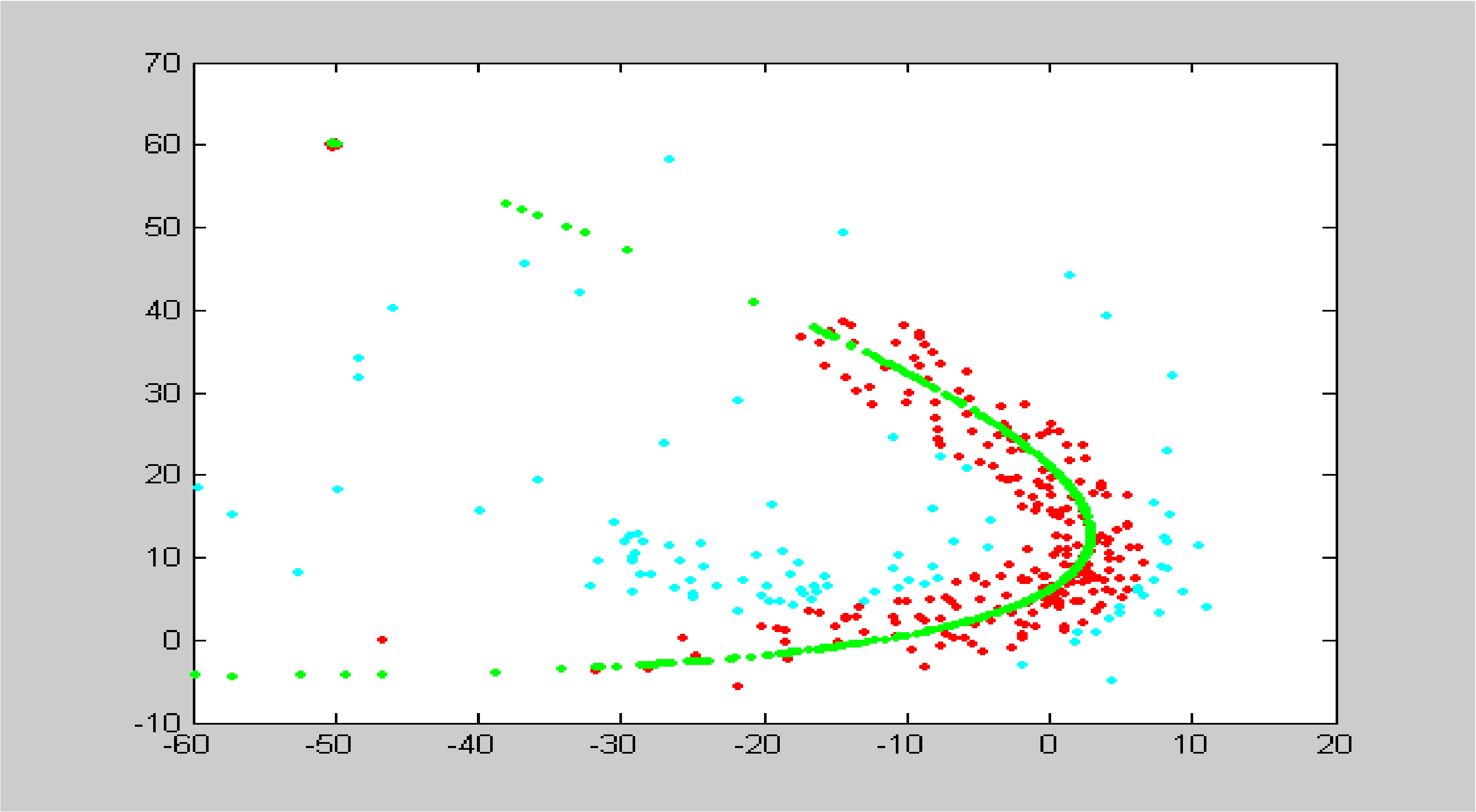


Clustering around parametrical curves

García-Escudero, M-I, Sánchez-Gutiérrez (CSDA, 2017)

Parabola

background noise=0.08 pointwise contamination=0.16 trimming level=0.25

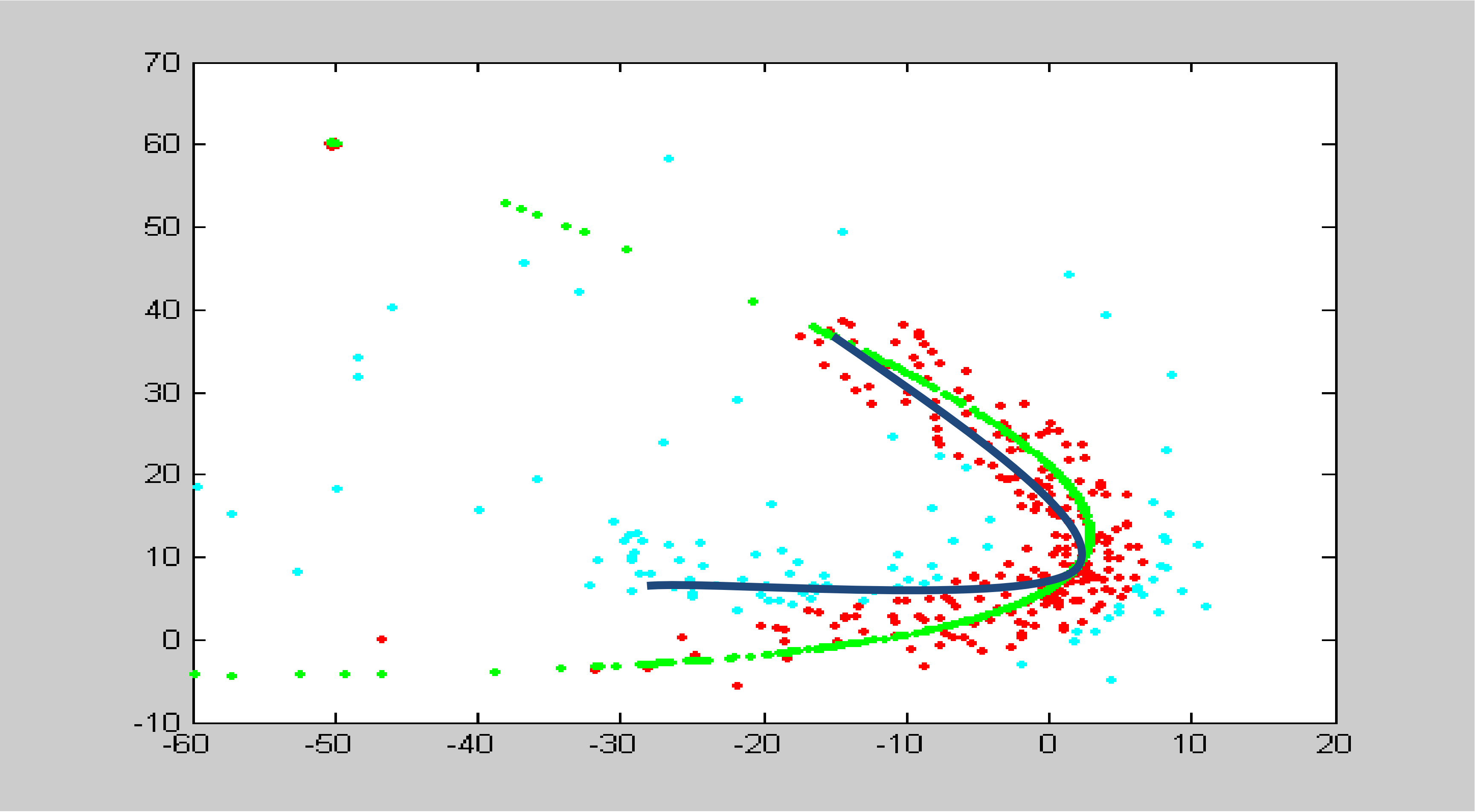


Clustering around parametrical curves

García-Escudero, M-I, Sánchez-Gutiérrez (CSDA, 2017)

Parabola

background noise=0.08 pointwise contamination=0.16 trimming level=0.25



Clustering around parametrical curves

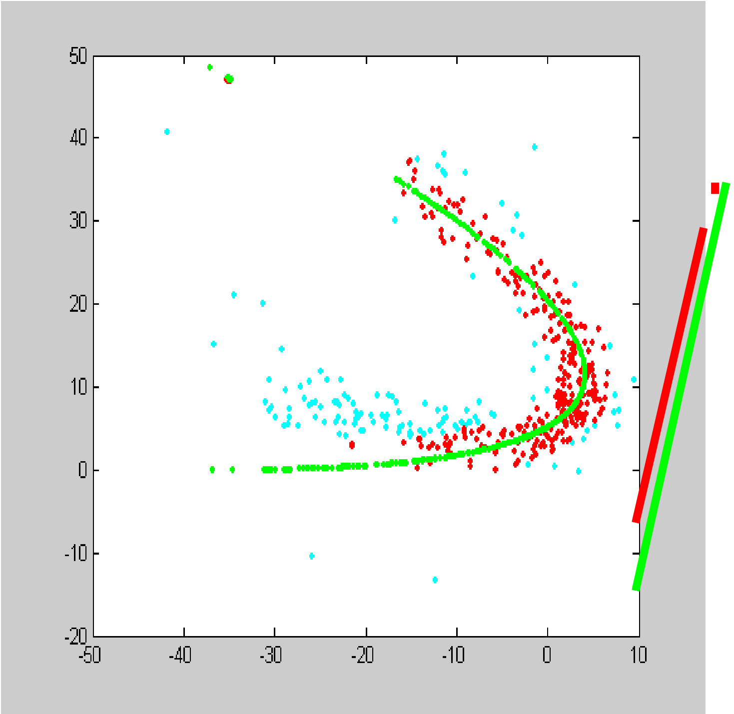
García-Escudero, M-I, Sánchez-Gutiérrez (CSDA, 2017)

Second trimming step

For avoiding the influence of pointwise contamination

This trimming is applied inside E step using the survival observations from the first trimming step

It is carried by using a α_2-trimmed mean over the proyected points in an orthogonal direction to the main axe of the parabola

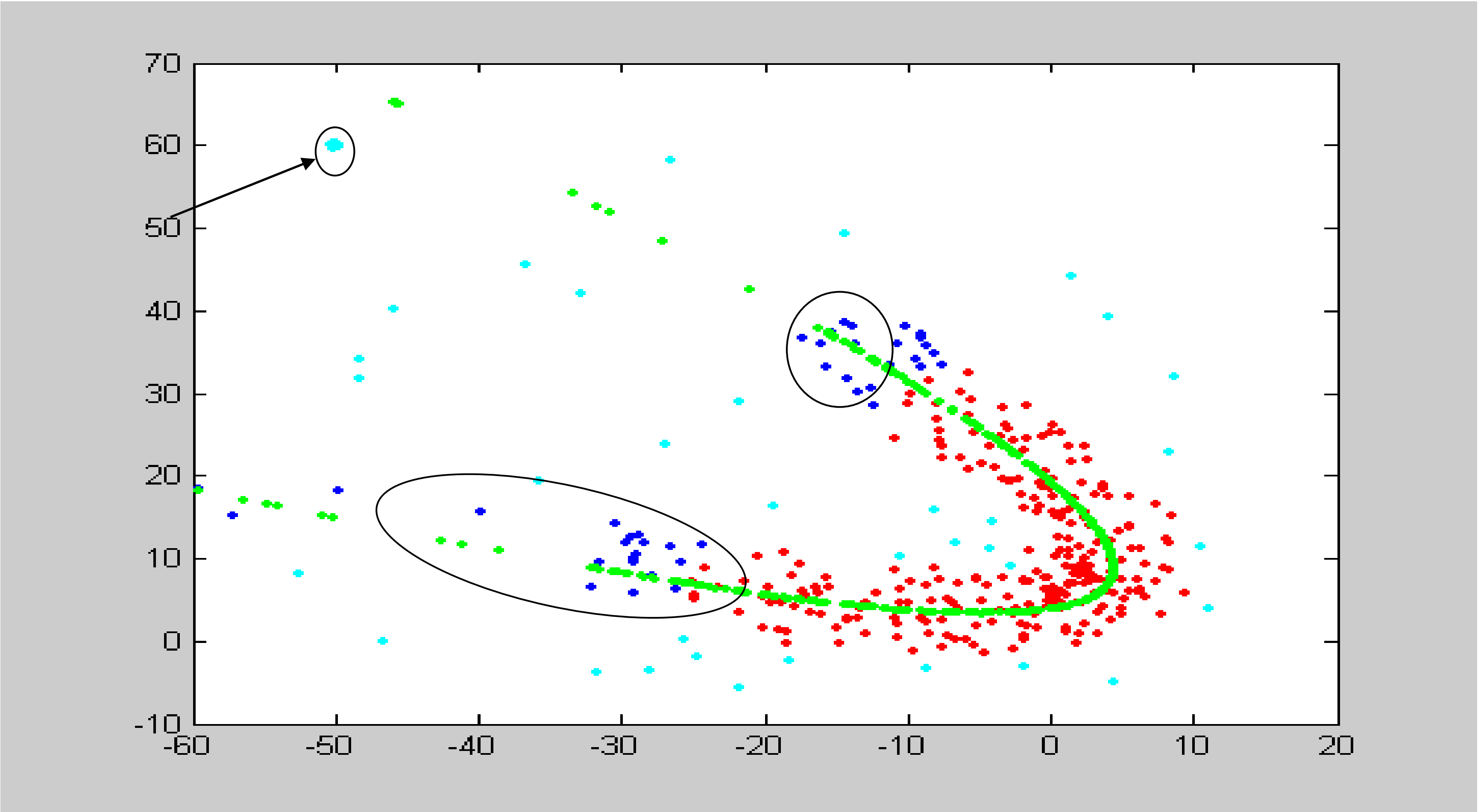


Clustering around parametrical curves

García-Escudero, M-I, Sánchez-Gutiérrez (CSDA, 2017)

Parabola

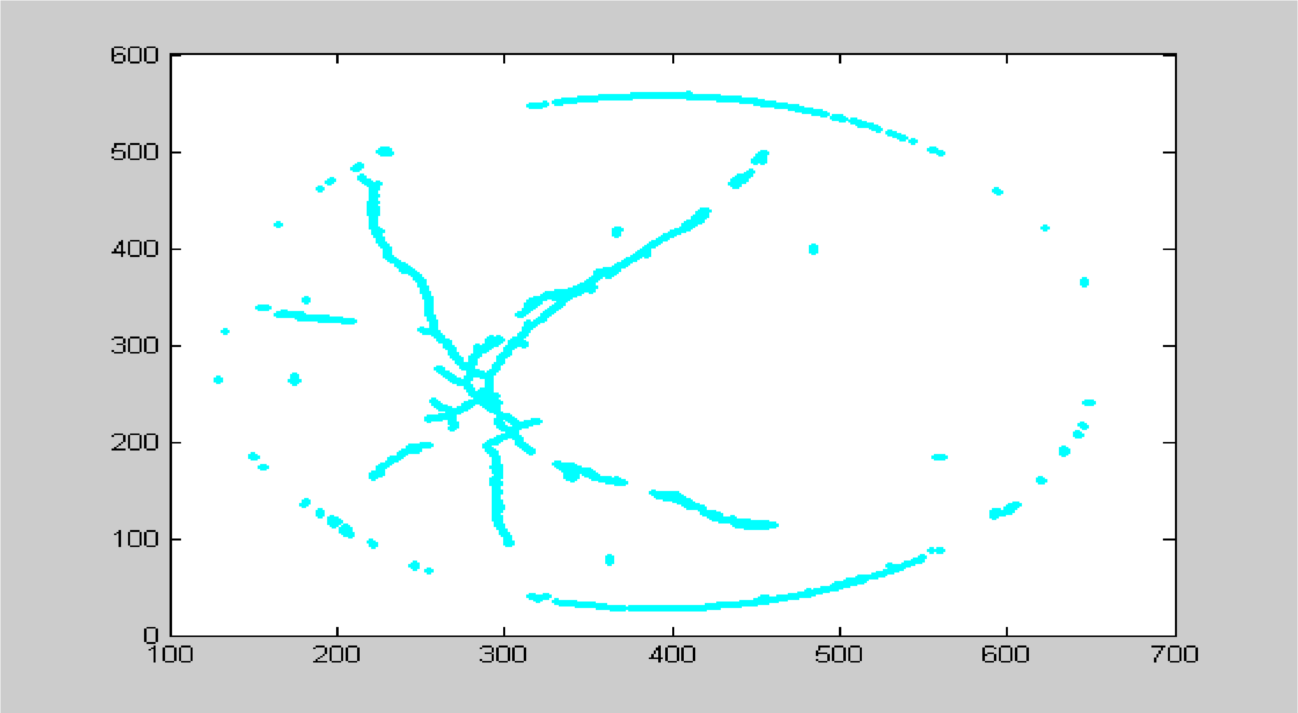
background noise=0.08 pointwise contamination=0.16
level of trimming1=0.25 trimming level=0.15



Clustering around parametrical curves

García-Escudero, M-I, Sánchez-Gutiérrez (CSDA, 2017)

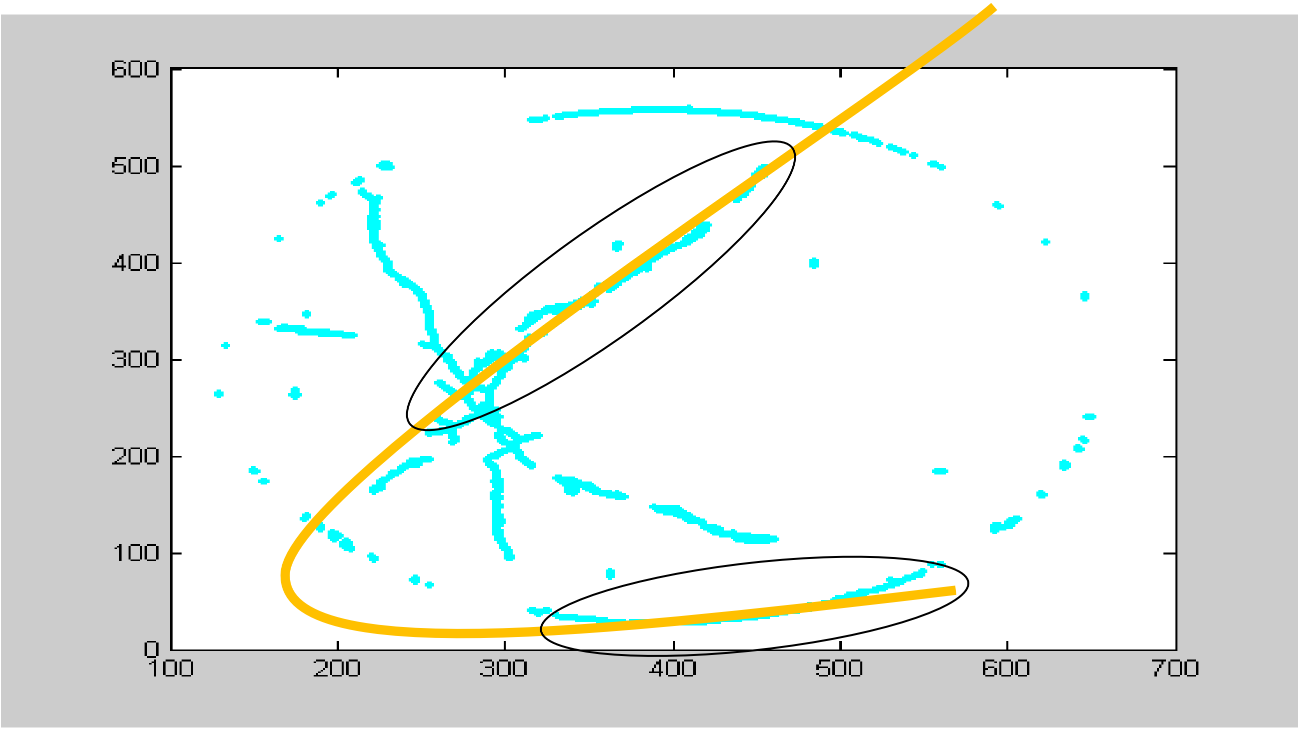
Retinography



Clustering around parametrical curves

García-Escudero, M-I, Sánchez-Gutiérrez (CSDA, 2017)

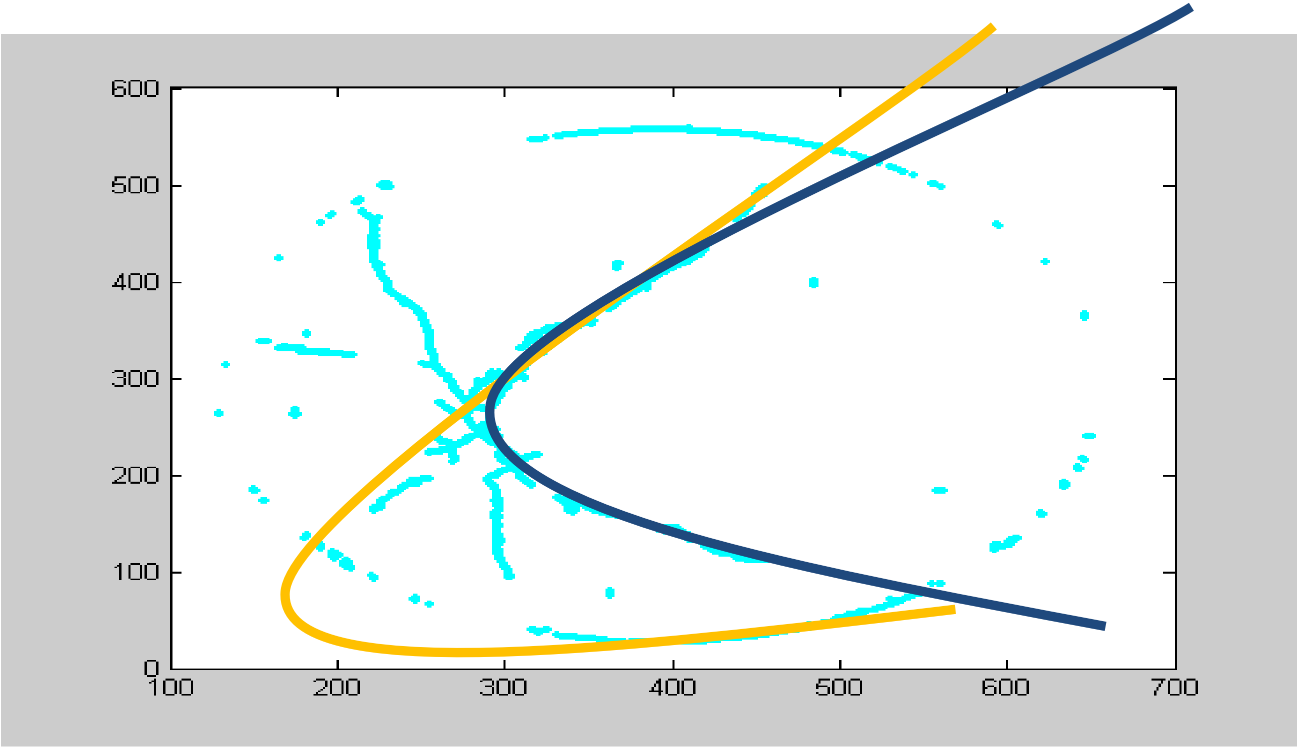
Retinography



Clustering around parametrical curves

García-Escudero, M-I, Sánchez-Gutiérrez (CSDA, 2017)

Retinography



Clustering around parametrical curves

García-Escudero, M-I, Sánchez-Gutiérrez (CSDA, 2017)

Retinography