Introduction to Robust Statistics

Anthony Atkinson, London School of Economics, UK

Marco Riani, Univ. of Parma, Italy

Multivariate analysis

Multivariate location and scatter

- Data $y_1, y_2, \ldots, y_n$ where the observations are v x 1 column vectors
- Classical model: elliptical distribution (e.g. multivariate gaussian) with parameters μ (location) and Σ (spread and correlation structure)

Remarks

- In the multivariate setting outliers cannot be detected by applying outlier detection rules to each variable separately
- There is no natural ordering of multivariate data

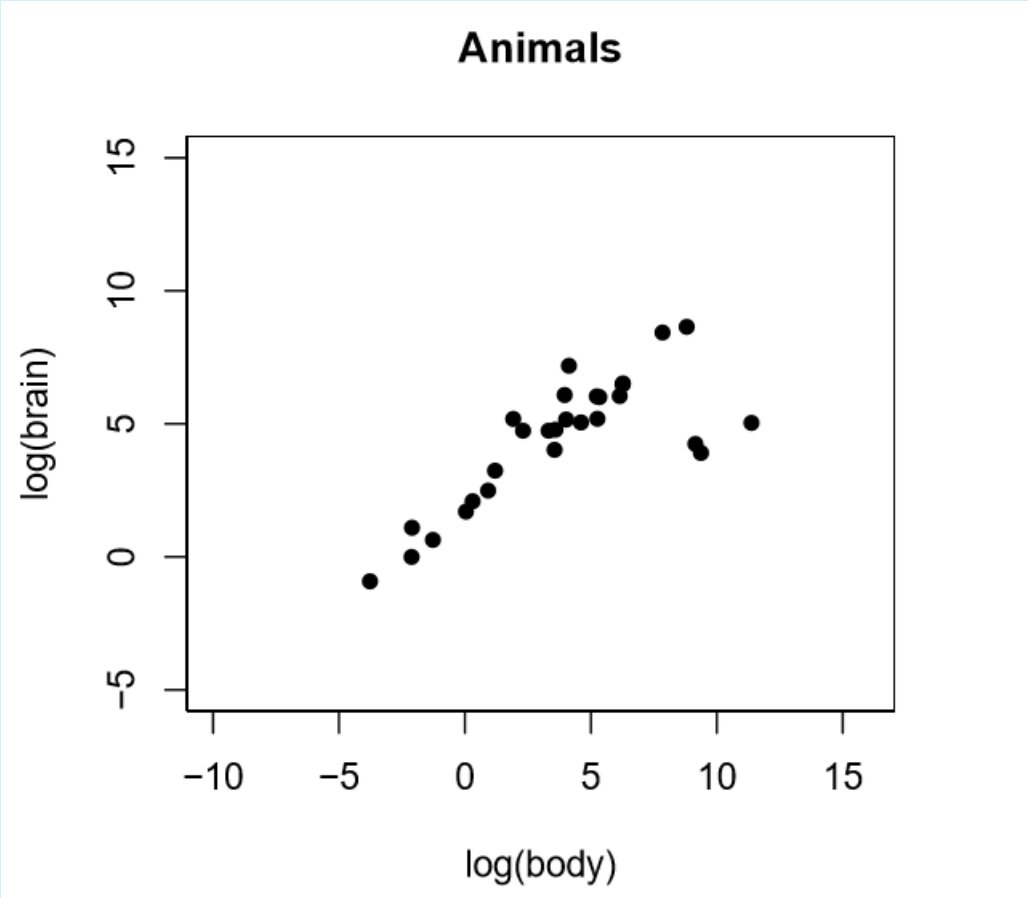
Target

- identify a large portion of the outliers when they are present in the data**: high power**
- provide a small number of false alarms with good data (i.e. data coming from the postulated model): **low swamping**

Bivariate data

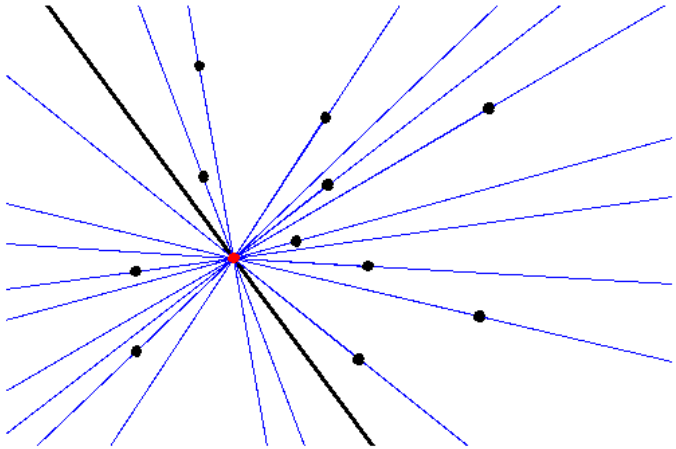
Ex. animals data set

- Consider the Animals data set containing the logarithm of the body and brain weight of 28 animals



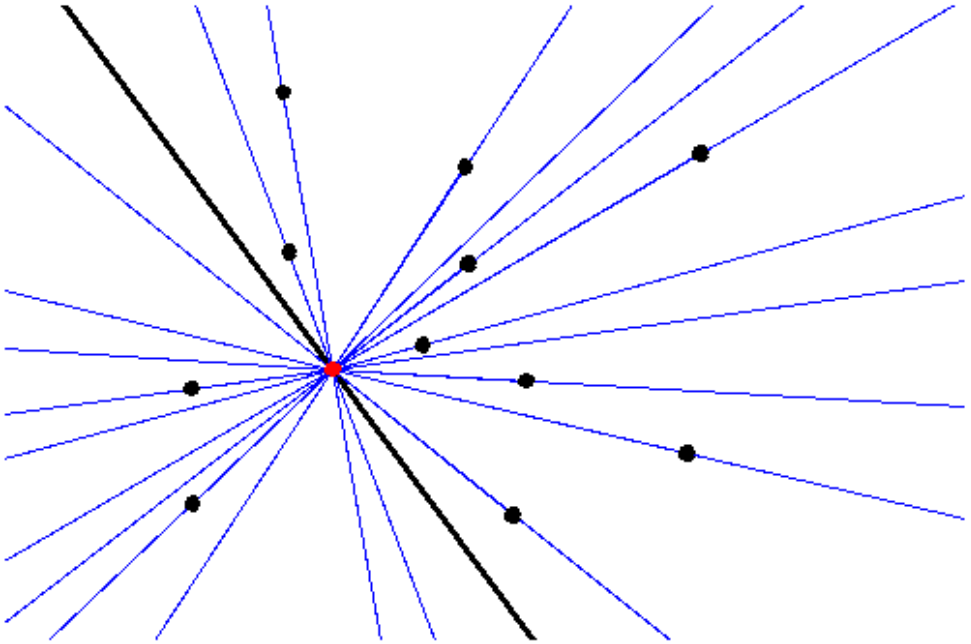
(Tukey) depth

- Depth is a generalization of ranking to multivariate situations. It is a nonparametric notion, since it is not assumed that the data come from a given type of distribution (e.g. elliptical).
- For bivariate data, the halfspace depth of a point y is the smallest number of observations in any halfplane whose boundary passes through y. Points on the outskirts have low depth, whereas points in the middle get high depth

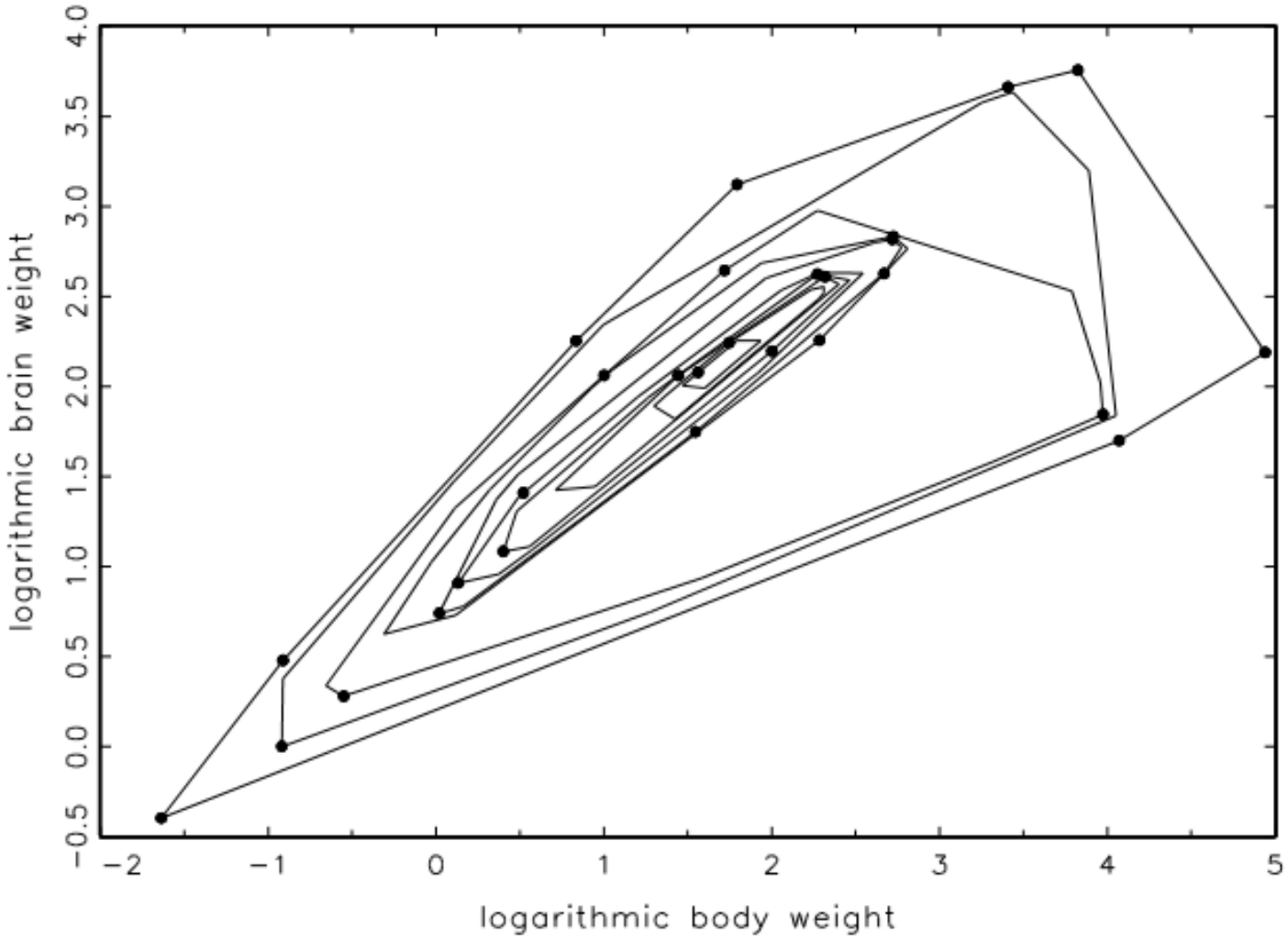


Depth in 2 dimensions

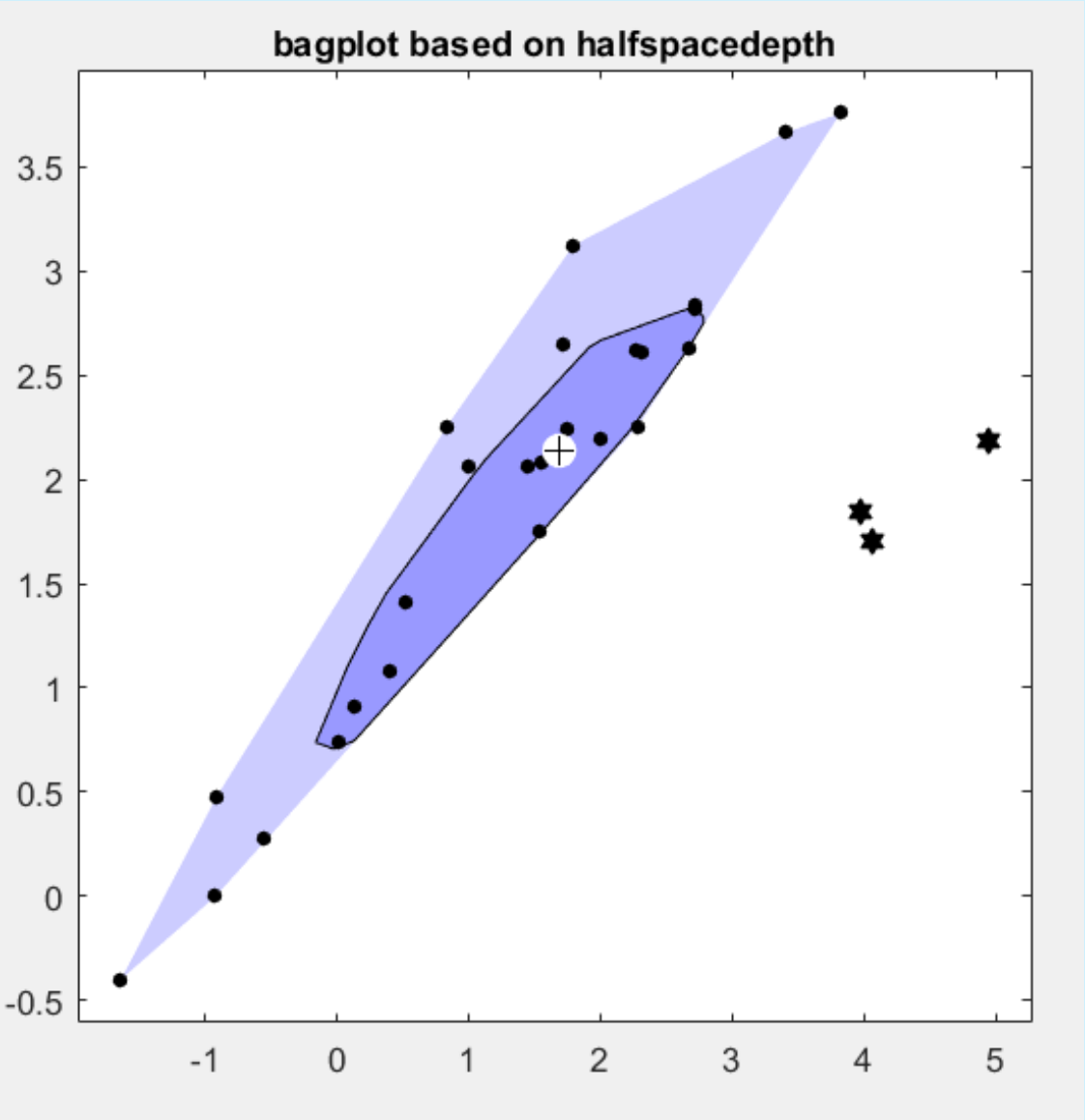
- For example, in the figure below the Tukey depth of the red point is 2 because the heavy line has two points on its left and every other line has at least two points on its left and right.



Example of depth contours

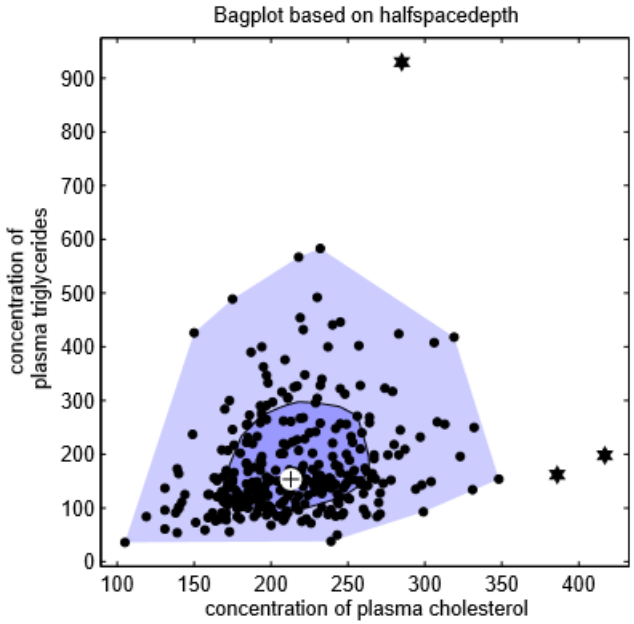


Bagplot of the animal data



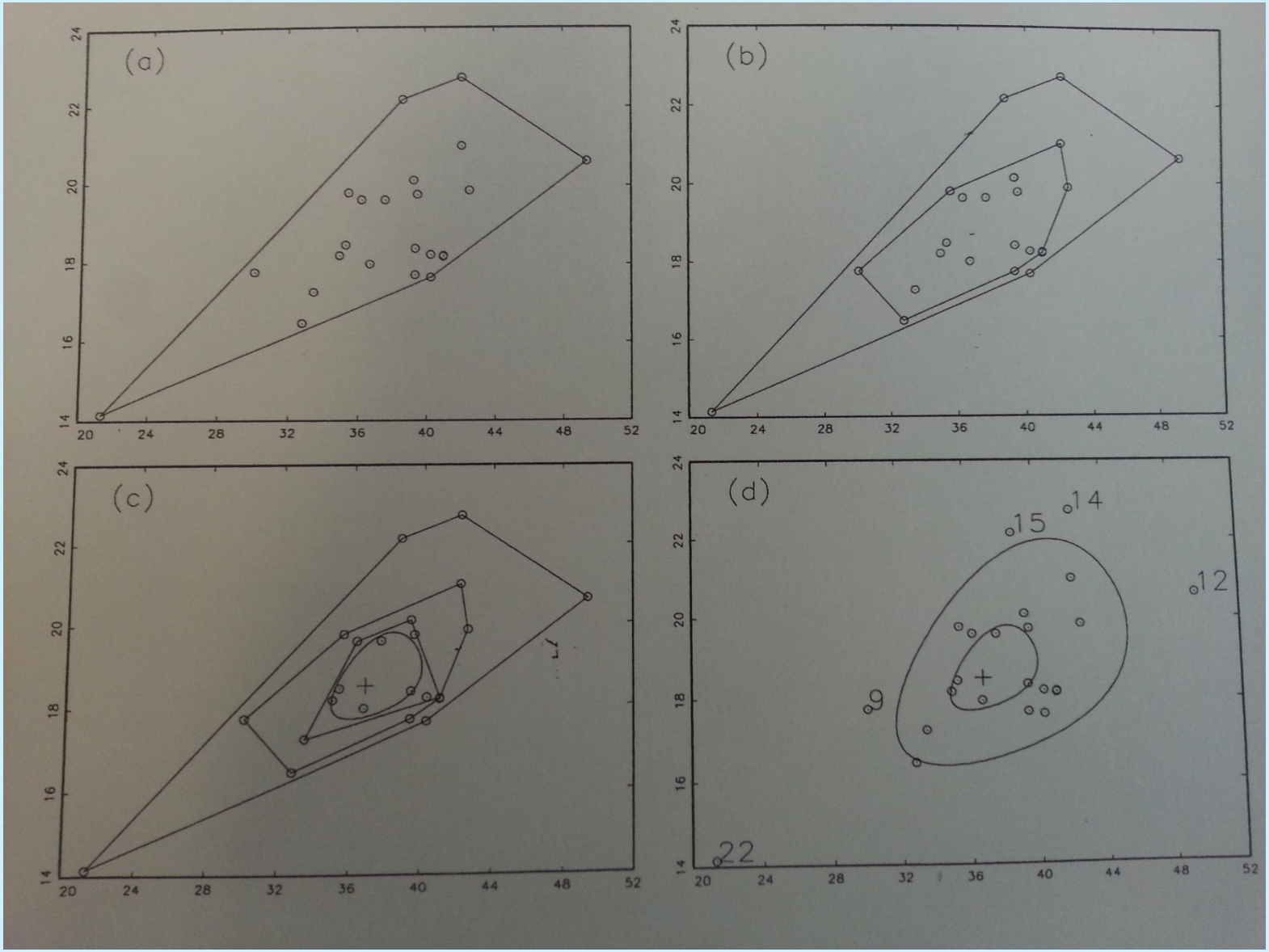
Bagplot

- The bagplot is a bivariate generalisation of the univariate boxplot

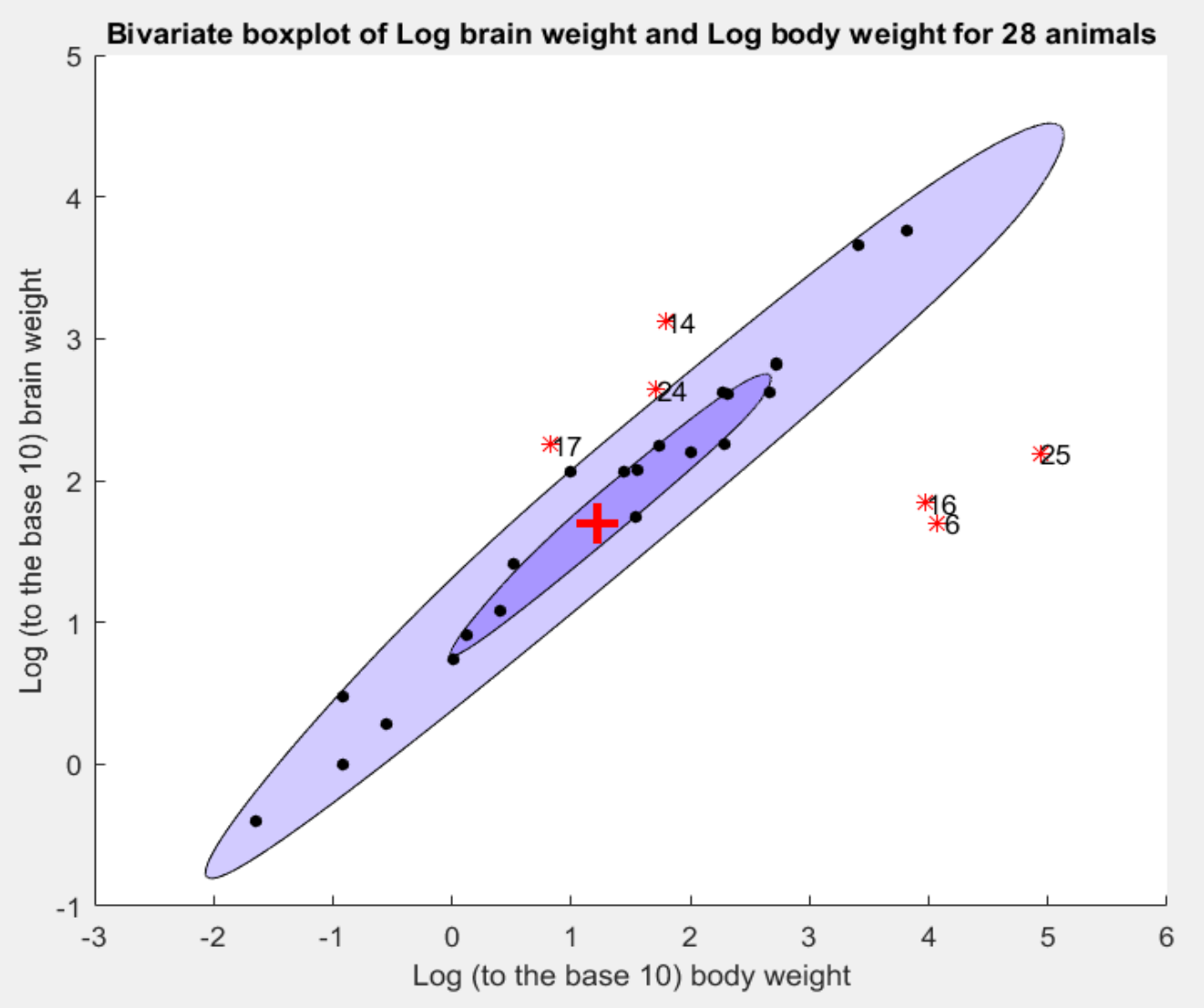


- The bagplot visualizes several characteristics of the data: its location, spread (the size of the bag), correlation (the orientation of the bag), skewness (the shape of the bag and the outer contour), and tails (the points near the boundary of the outer contour and the outliers).

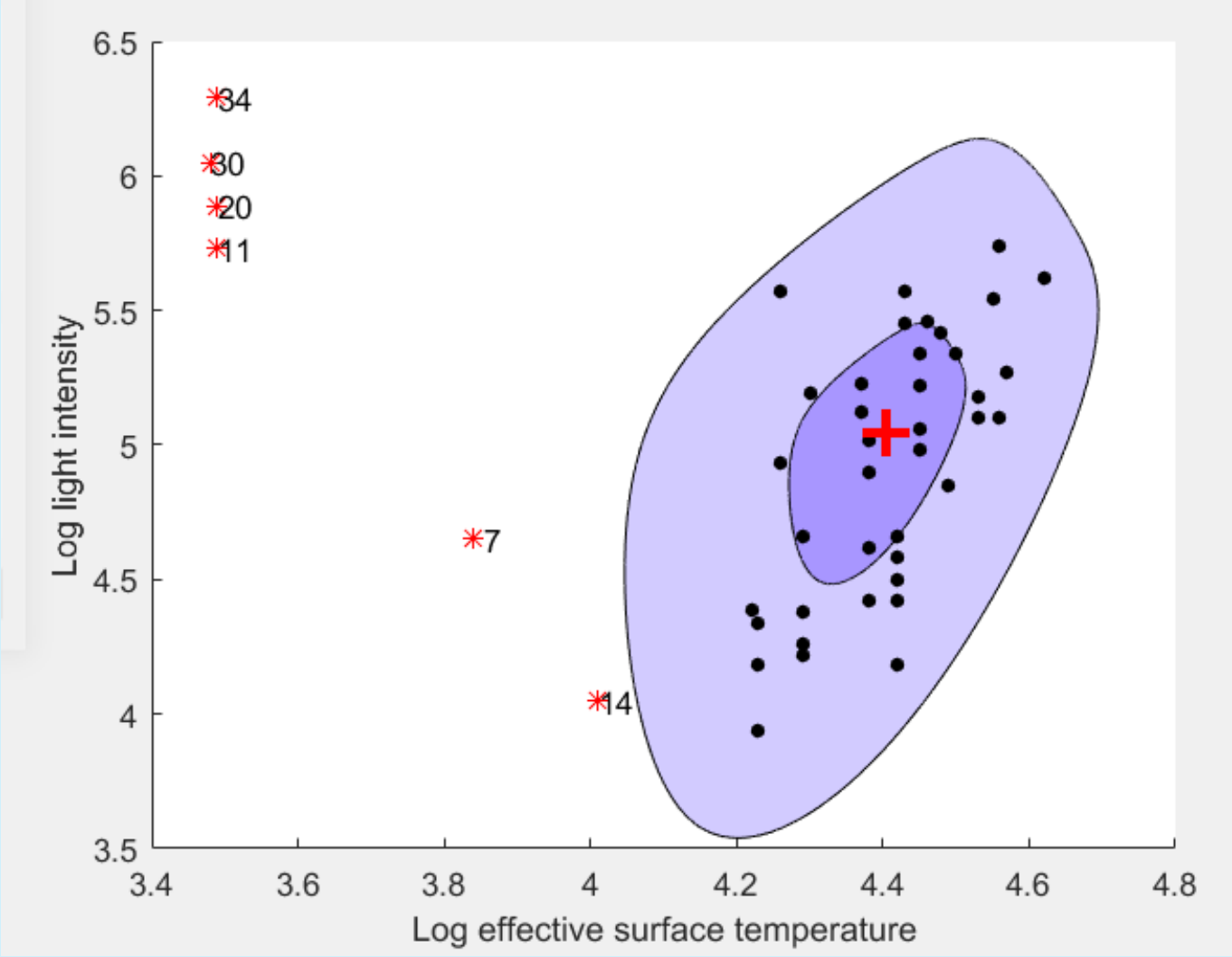
Bivariate boxplot based on convex hull peeling



Bivariate boxplot of the animals data



Bivariate boxplot of the stars data



Multivariate data

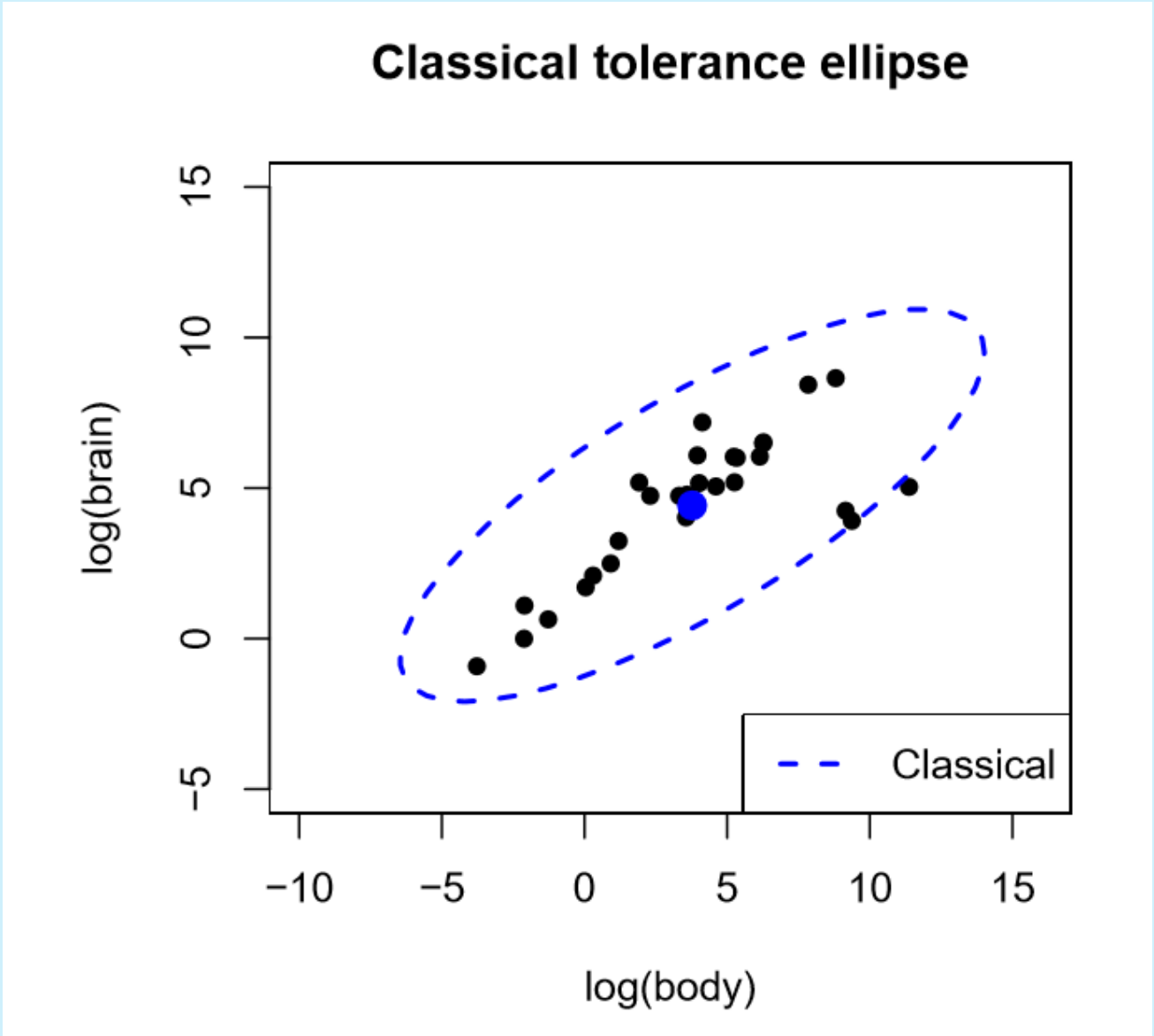
Mahalanobis distances

- If we estimate the parameters of $N(\mu, \Sigma)$ with $\hat{\mu}$ sample mean and $\hat{\Sigma}$ unbiased sample covariance matrix multivariate outliers should have large **Mahalanobis Distances** (MD):
- $d_i^2 = (y_i - \hat{\mu})^T \hat{\Sigma}^{-1} (y_i - \hat{\mu})$
- For statistical outlier detection, we thus need **cut-off values** for d^2_i We can also assume that $\mu = \mathbf{0}$ and $\mathbf{\Sigma} = \mathbf{I}$ because MD are invariant under **affine transformations**

Tolerance ellipsoid

- Its boundary contains those y-values with constant MD to the mean
- Classical tolerance ellipsoid
- $\{y; MD(y) \leq \sqrt{\mathrm{X}^2_{v,0.975}}\}$
- with $\mathrm{X}^2_{v,0.975}$ the 97.5% quantile of the X^2 distirbution with v degrees of freedom
- We expect (for large n) that about 97.5% of the observations belong to the ellipsoid

(Classical) tolerance ellipsoid



Scatter ratios

- Wilks showed that under the **null hypothesis of no outliers**
- $H_0: \{y_1 \sim N(\mu,\Sigma)\} \cap \{y_2 \sim N(\mu,\Sigma)\} \cap \ldots \cap \{y_n \sim N(\mu,\Sigma)\}$
- the n **scatter ratios**

$$R_i = |(n-2)\hat{\Sigma}_{\{i\}}||(n-1)\hat{\Sigma}|^{-1} = 1-\frac{n}{(n-1)^2}d_i^2 \qquad i = 1,\ldots,n,$$

- have a $Beta\left(\frac{n-v-1}{2},\frac{v}{2}\right)$ distribution

Outlier test

- A **Bonferroni bound** can be used to approximate the distribution of the smallest ratio R_1 or equivalently of the largest squared distance $d^2_{(n)}$

- Test for outlyingness of the most extreme observations

Wilks rule

- Compute the largest squared distance $d^2_{(n)}$
- **At level** γ, label the corresponding observation an **outlier** if

$$d^2_{(n)} > \frac{(n-1)^2}{n} b_{1-\gamma/n}$$

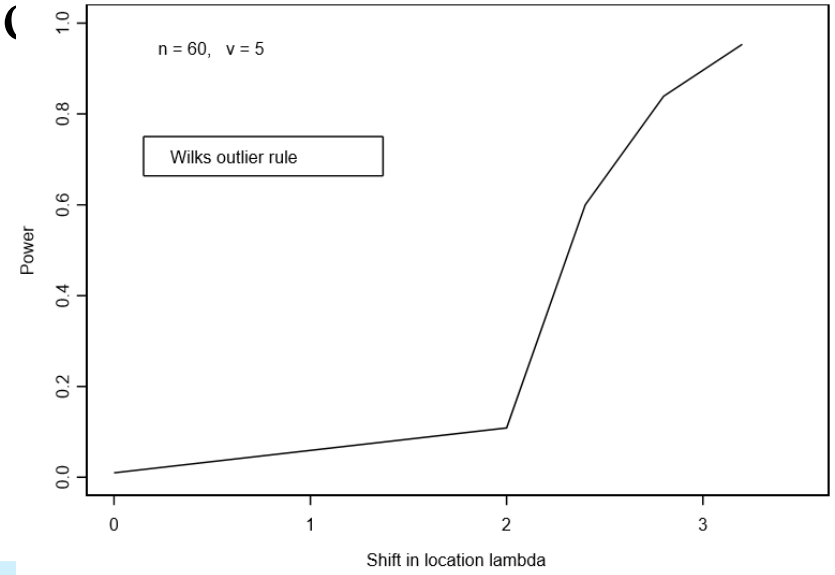
- where $b_{1-\gamma/n}$ is the $1-\gamma/n$ quantile of the $Beta\left(\frac{n-v-1}{2}, \frac{v}{2}\right)$ distribution

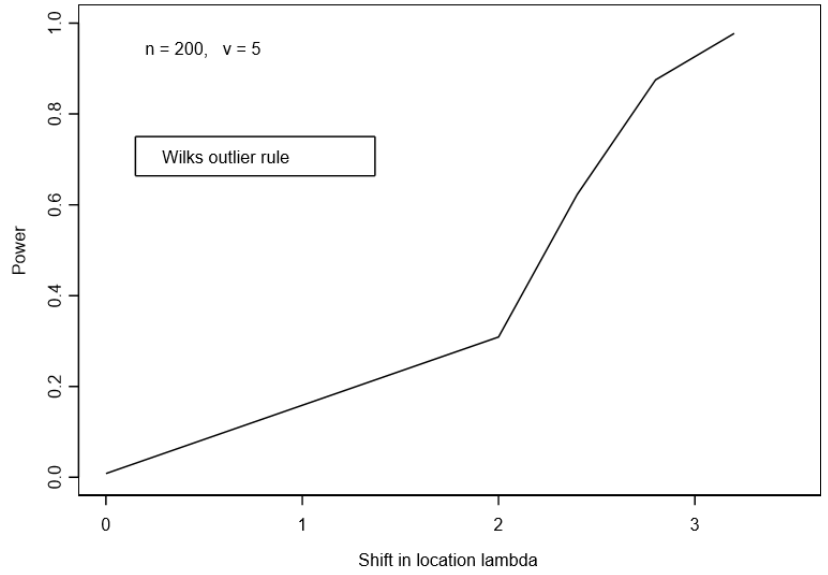
Wilks rule - simulations

- 5,000 simulations for each combination of n and v.
- **Size** under the null model of no contamination: $N(0.I)$

	$n = 40$	$n = 90$	$n = 200$	$n = 400$
$v = 5$	0.010	0.009	0.008	0.008
$v = 10$	0.010	0.009	0.012	0.009
$v = 15$	0.009	0.011	0.012	0.009

- **Power** under a location-shift contamination model for one

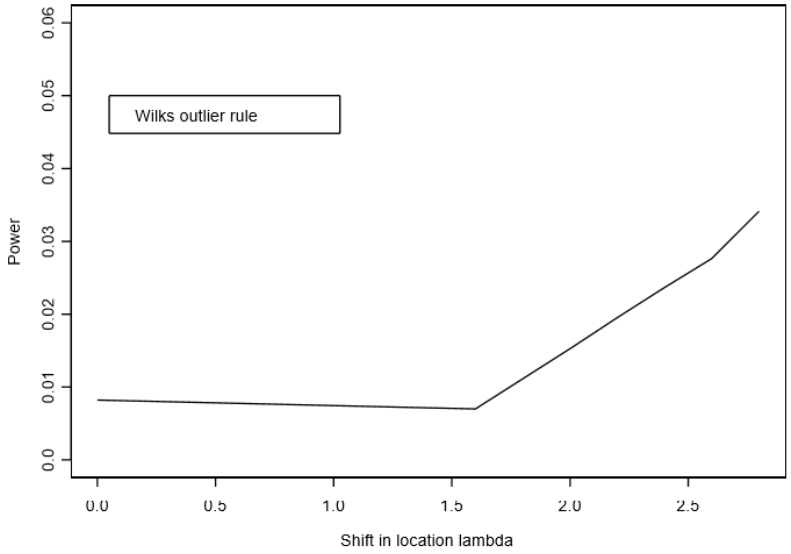


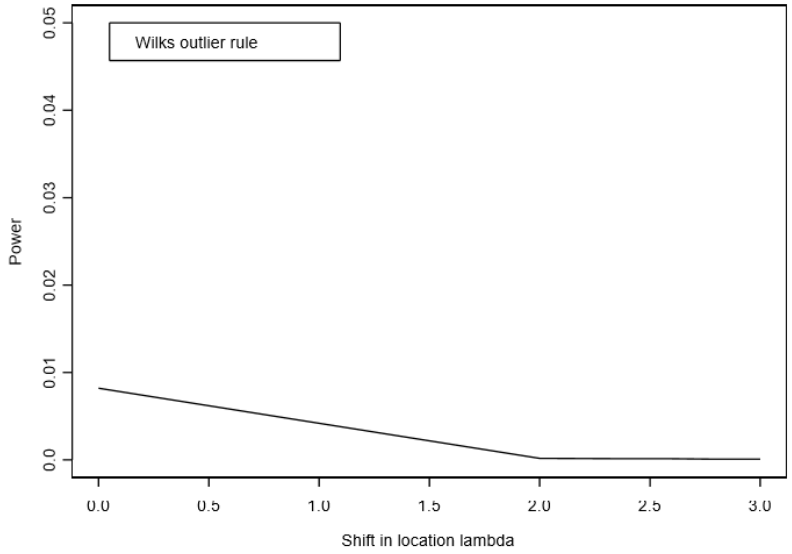


Wilks rule - summary

- The Wilks rule is a statistically principled criterion for multivariate outlier detection with good properties both in small and large samples: •
- the probability of detecting contamination increases with the amount of contamination **when the data contain a single outlier**
- this probability approaches 1 if λ is large enough
- the actual size of the test is very close to the nominal γ **when no outlier is present**
- These goals are achieved through:
- **accurate distributional results**: Beta distribution
- recognition that we perform n **simultaneous tests**: compute the largest squared Mahalanobis distance
- **But ...**

- **With more than 1 outlier, what is the proportion of outliers detected by the Wilks rule?**
- Nominal size: $\gamma = 0.01$, $n = 200$, $v = 5$. 5,000 simulations for each λ.
- **Left panel: 5% of the observations are contaminated**
- **Right panel: 20% of the observations are contaminated**

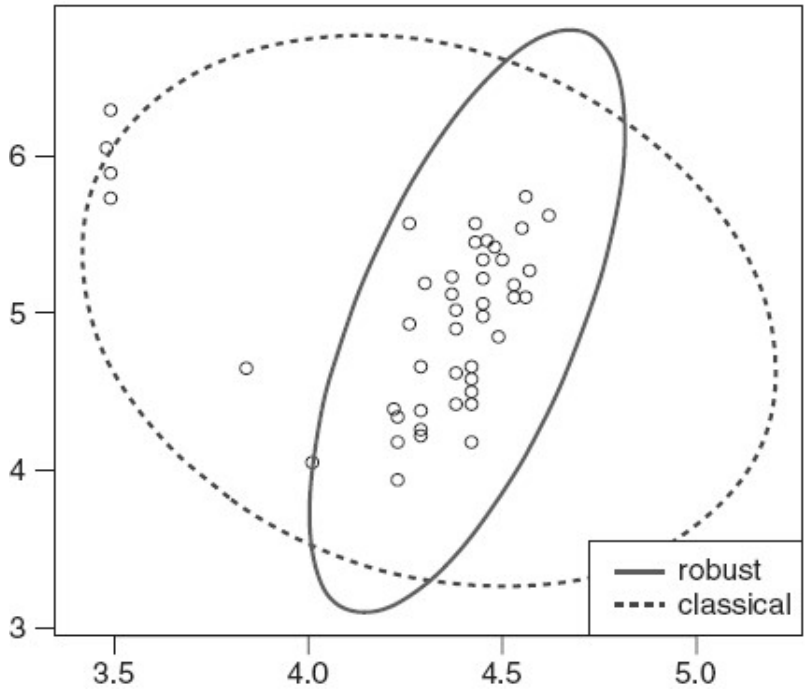




- **Wilks rule is ineffective**: this phenomenon is known as **masking**

Masking

- Masking occurs because the classical estimates $\hat{\mu}$ (sample mean) and $\hat{\Sigma}$ (unbiased sample covariance matrix) **are grossly biased by the presence of many (extreme) outliers.**
- **Ex. Stars data**



- The ellipses represent 0.99 probability contours. The large ellipse is based on $\hat{\mu}$ and $\hat{\Sigma}$. The small ellipse is obtained by computing **robust high-breakdown estimates** of μ and Σ!

Breakdown point

- In regression we used the expression: «bounded and also bounded away from the boundary of the parameter space»
- Dispersion matrices: parameter space consists of the set of symmetric non negative definite matrices
- Each matrix has eigenvalues and eigenvectors.
- $\hat{\Sigma}$ «bounded and also bounded away from the boundary of the parameter space» means: the eigenvalues are bounded away from 0 and infinity

Efficiency

- All affine equivariant location estimates $\hat{\mu}$ (when data are from $N(\mu, \Sigma)$) have an asymptotic covariance matrix of the form $\gamma\Sigma$ where γ is a constant depending on the estimate.

- Consequence: the normal distribution efficiency of an affine equivariant location estimate is independent of μ and Σ

Multivariate M estimators

- The multivariate M-estimate of location and dispersion are defined as the solution of the following system of equations (estimating equations)

$$\begin{cases} \sum_{i=1}^{n} W_1(d_i)(y_i - \hat{\mu}) = 0 \\ \frac{1}{n}\sum_{i=1}^{n} W_2(d_i)(y_i - \hat{\mu})(y_i - \hat{\mu})' = \hat{\Sigma} \end{cases}$$

where the functions $W1$ and $W2$ need not to be equal

- If function *W*2 is non decreasing, the solution to this system of equation is called monotone multivariate M estimates, while if *W*2 is redescending the solutions are called redescending multivariate M-estimates

Properties of multivariate M estimators

- Note that from

$$\begin{cases} \sum_{i=1}^{n} W_1(d_i)(y_i - \hat{\mu}) = 0 \\ \frac{1}{n}\sum_{i=1}^{n} W_2(d_i)(y_i - \hat{\mu})(y_i - \hat{\mu})' = \hat{\Sigma} \end{cases}$$

- we can express $\hat{\mu}$ as a weighted mean, with weights depending on the outlying measure d_i (weighted mean with data dependent weights)
- $\hat{\mu} = \sum_{i=1}^{n} W_1(d_i) y_i / \sum_{i=1}^{n} W_1(d_i)$

- Multivariate M estimates are affine equivariant and asymptotically have a multivariate normal distribution.

Numerical computations of multivariate estimates

- $$\begin{cases} \sum_{i=1}^{n} W_1(d_i)(y_i - \hat{\mu}) = 0 \\ \frac{1}{n}\sum_{i=1}^{n} W_2(d_i)(y_i - \hat{\mu})(y_i - \hat{\mu})' = \hat{\Sigma} \end{cases}$$

- Start with initial estimates $\hat{\mu}_0$ and $\hat{\Sigma}_0$ (vector of coordinate-wise medians and the diagonal matrix with the squared normalized MADs of the variables in the diagonal). At iteration k let $d_{k,i} = d(y_i, \hat{\mu}_k, \hat{\Sigma}_k)$ and compute
- $\hat{\mu}_{k+1} = \sum_{i=1}^{n} W_1(d_{k,i}) y_i / \sum_{i=1}^{n} W_1(d_{k,i})$
- $\hat{\Sigma}_{k+1} = \sum_{i=1}^{n} W_2(d_{k,i})\ (y_i - \hat{\mu}_{k+1})\ (y_i - \hat{\mu}_{k+1})^T$

M and S estimators

- Just as with the regression estimates where we aimed at making the residuals "small", we shall define multivariate estimates of location and dispersion that make the distances d_i small. To this purpose we look for $\hat{\mu}$ and $\hat{\Sigma}$ some measure of largeness of $d^2(y, \hat{\mu}, \hat{\Sigma})$.
- Avoid spurious solutions: exclude solutions for which the smallest eigenvalue of $\hat{\Sigma}$ is zero
- If $\hat{\Sigma} = \sigma^2|\hat{\Gamma}|$ impose constraint $|\hat{\Gamma}|=1$

M and S estimators

- Minimize a robust estimate of scale

$$\min_{\hat{\mu} \in R^v,\, \hat{\Gamma} \in S_v \text{ with } |\hat{\Gamma}|=1} \hat{\sigma}\left(d(y, \hat{\mu}, \hat{\Gamma})\right)$$

- If $\hat{\sigma}$ is an M scale estimate which satisfies

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{d_i}{\hat{\sigma}}\right) = K,$$

- where ρ is a smooth bounded ρ-function, we obtain the class of S estimates. *K* for consistency is chosen as

$$K = E_{\Phi}[\rho(d(y, 0, I))]$$

Characteristics of S estimates of multivariate location and scatter

- Affine equivariant
- BDP $\epsilon^* \approx \min\{K/\rho(\infty), 1 - K/\rho(\infty)\} \leq 50\%$
- Bounded influence function
- (Fisher)-consistent and asymptotically normal
- Constant c in Tukey biweight controls *bdp* (and *eff*)
- Efficiency is low

MM estimators

- The MM estimator of location and shape is defined as the minimum of the following *f* function

$$f(\hat{\mu}_{MM}, \hat{\Gamma}_{MM}) = \min_{\mu \in R^v,\, \Gamma \in S_v \text{ with } |\Gamma|=1} \frac{1}{n} \sum_{i=1}^{n} \rho_2 \left(\frac{d(y_i, \mu, \Gamma)}{\hat{\sigma}} \right)$$

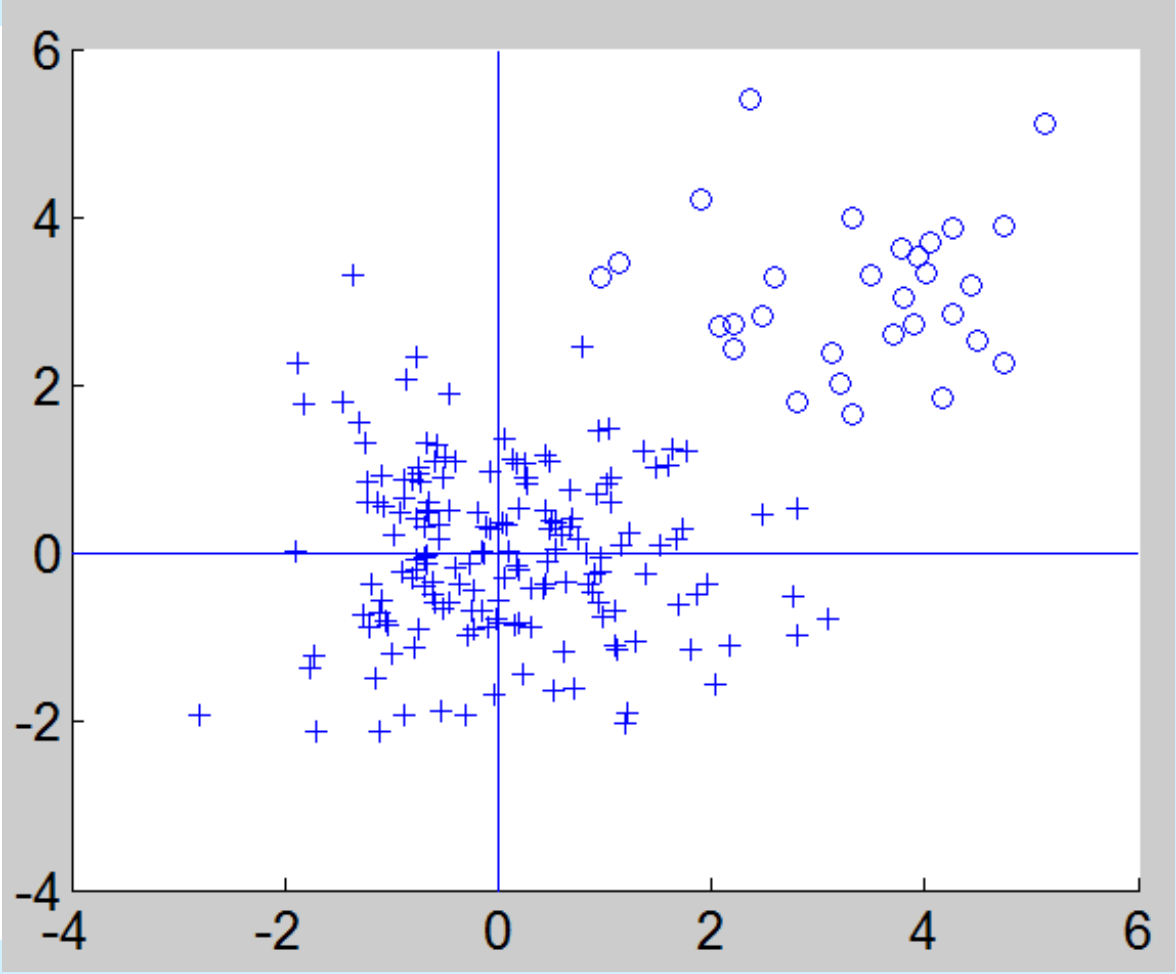
- where ρ_2 is possibly another ρ function which provides higher efficiency than the previous ρ at the null multivariate normal model. Function *f* is minimized with respect to μ and Σ for fixed $\hat{\sigma}$.
- $\hat{\sigma}$ is any auxiliary robust scale estimate, however it is common to use $\hat{\sigma}_S$ and as starting values of location and shape, those which come out from the S estimator (that is $\hat{\mu}_s$ and $\hat{\Sigma}_s$)

MM estimators

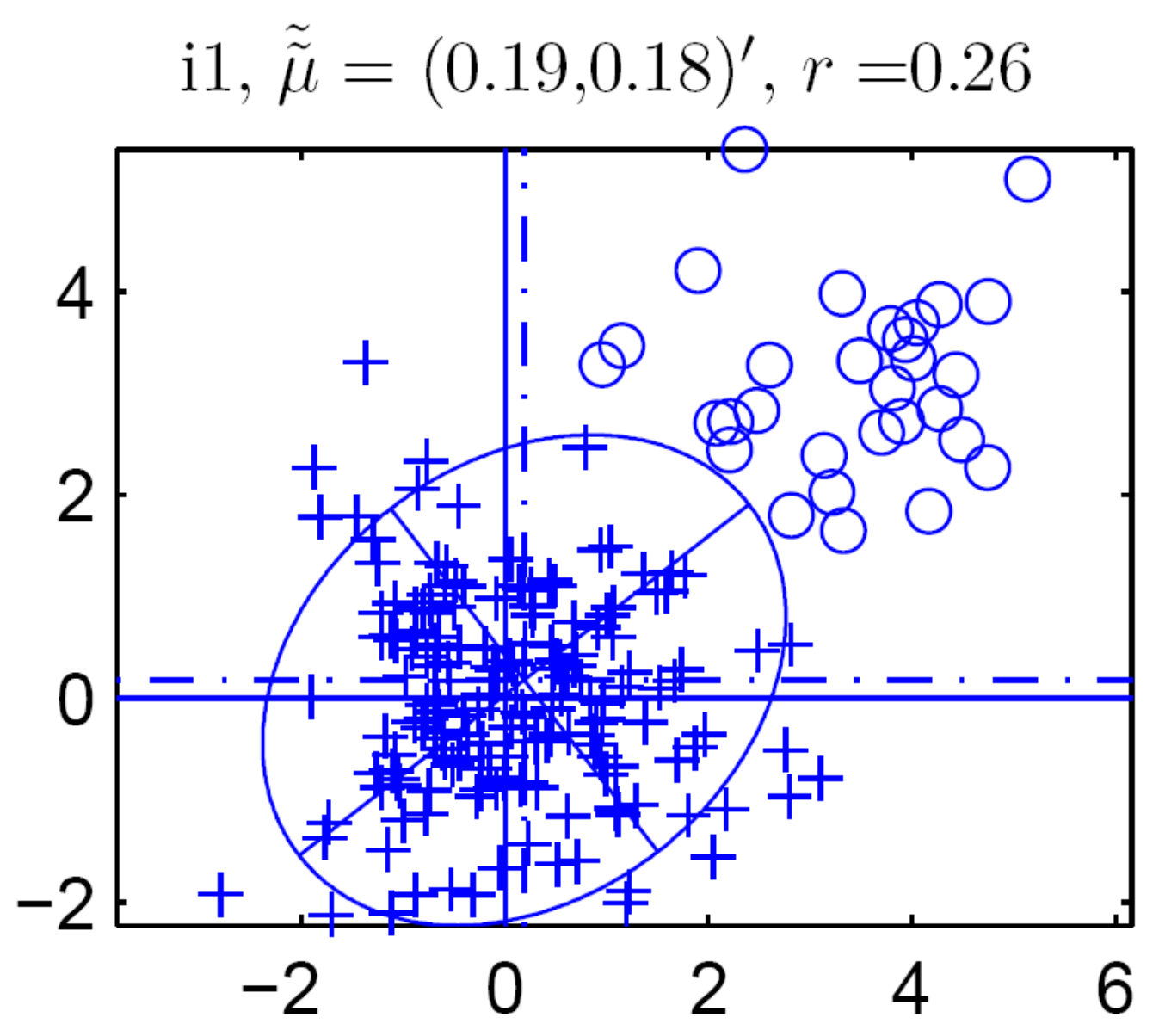
- The MM estimate of scatter is given by $\hat{\Sigma}_{\mathrm{MM}} = \hat{\sigma}_S \hat{\Gamma}_{MM}$
- S estimator of scale $\rho_{bdp=0.5}$ is tuned for robustness (high bdp)
- Redescending M-estimator $\rho_{eff=0.95}$ is tuned for high efficiency
- Claim: highly robust and efficient!

Explanation for the failure of multivariate MM estimators

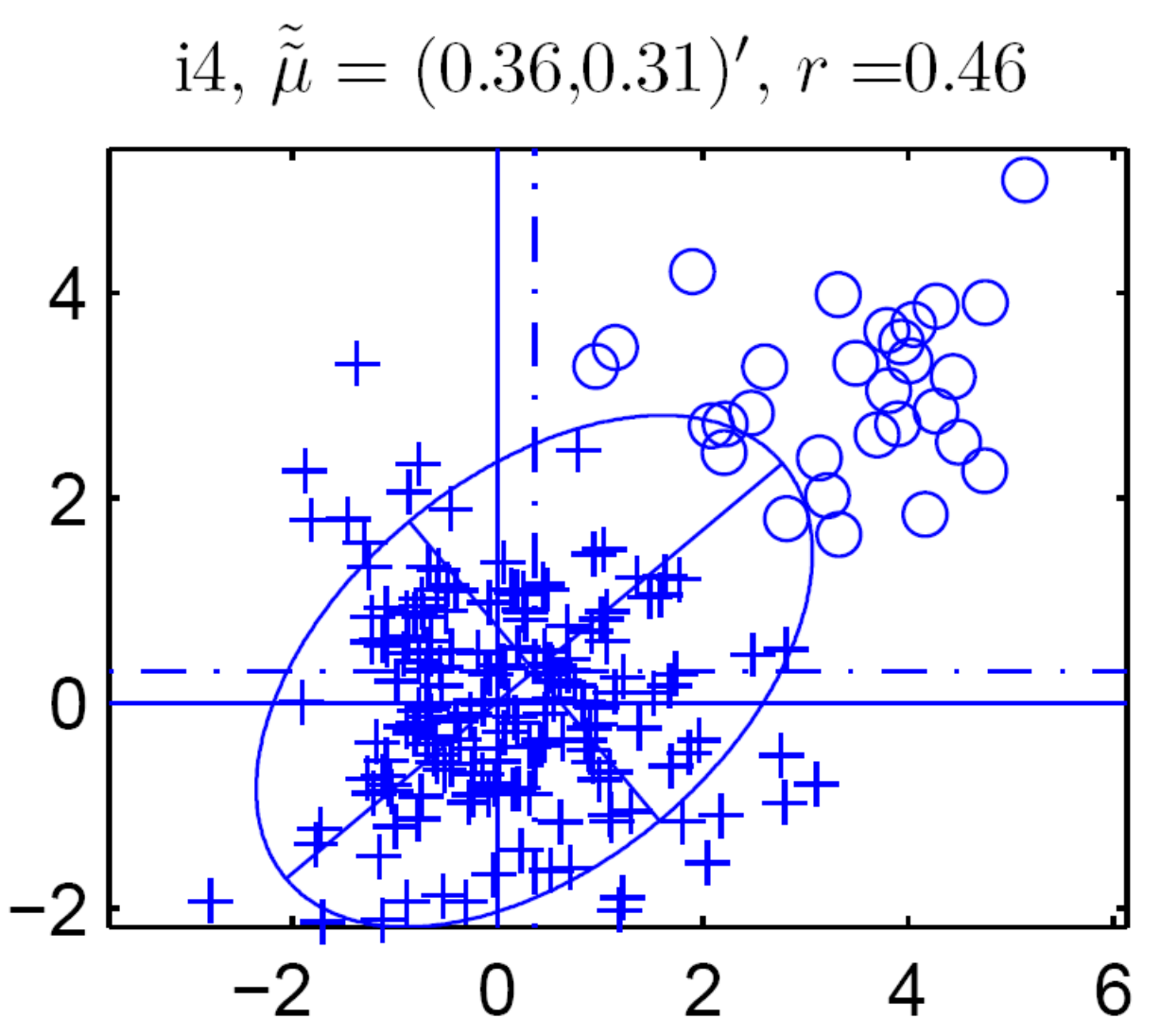
- n=200
- v=2
- δ=0.30
- shift of 3



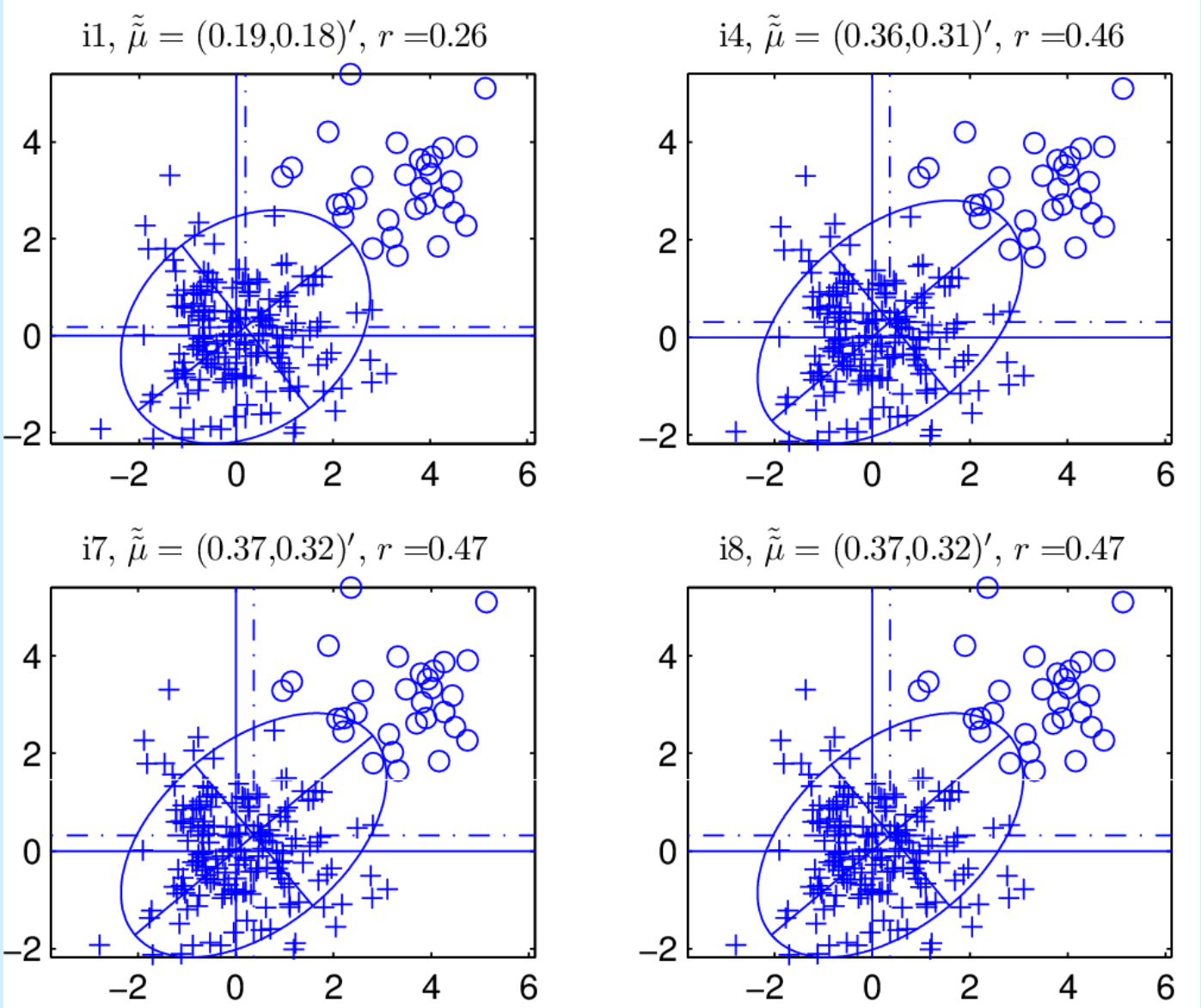
Iteration 1 in the MM loop



Iteration 4 in the MM loop



Iteration 1, 4, 7, 8 in the MM loop



MVE estimate

- If we take $\hat{\sigma}$ (to mimic the approach that results in the LMS in regression) the sample median of the Mahalanobis distances, the resulting location and dispersion matrix estimate is called minimum volume ellipsoid (MVE) estimate
- The name comes from the fact that among all ellipsoids $\{y: d(y, \mu, \Sigma)\}$ containing at least half of the data points, the one given by the MVE estimate has minimum volume.
- The consistency rate of MVE is the same slow rate as the LMS (namely $n^{-1/3}$)

MCD (Minimum covariance determinant)

- Idea: use a trimmed scale for $\hat{\sigma}$ instead of an M-scale (as was done to obtain LTS)
- More formally, let

$$d_{(1)}(\mu, \Gamma) \leq d_{(2)}(\mu, \Gamma) \leq \ldots \leq d_{(n)}(\mu, \Gamma)$$

- be the ordered of the squared distances $d^2(y_i, \hat{\mu}, \hat{\Gamma})$ and for $1 \leq h < n$ define the trimmed scale as

$$\hat{\sigma} = \sum_{i=1}^{h} d_{(i)}$$

MCD (Minimum covariance determinant)

- $y_{(\mathrm{MCD})}$:= sub-sample of $n/2 \leq h < n$ observations whose covariance matrix has the smallest determinant

$$\hat{\mu}_{(\mathrm{MCD})} = \frac{1}{h} \sum y_i$$

$$\hat{\Sigma}_{(\mathrm{MCD})} = \frac{k_{\mathrm{MCD}}(h,n,v)}{h-1} \sum_{i \in y_{(\mathrm{MCD})}} (y_i - \hat{\mu}_{(\mathrm{MCD})})(y_i - \hat{\mu}_{(\mathrm{MCD})})'$$

- The proportionality term $k_{\mathrm{MCD}}(h,n,v)$ is crucial to ensure **consistency** and (approximate) **unbiasedness** of $\hat{\Sigma}_{(\mathrm{MCD})}$
- The proportionality term is formed by an analytic component (for consistency) and a simulation-based component (for unbiasedness)
- The coverage h must be fixed: usually $h \approx 0.5n$ $h \approx 0.75n$ yielding a breakdown value of 50% and 25% respectively.

REWEIGHTED MCD

- Reweighted subsample: give weight $w_i = 0$ to observations for which $d_{iMCD} > \chi^2_{v,0.975}$ and weight 1 otherwise
- Claim: improve efficiency while mantaining the same bdp

$$\hat{\mu}_{\text{RMCD}} = \frac{1}{m}\sum_{i=1}^{n} w_i y_i \qquad \text{with} \qquad m = \sum_{i=1}^{n} w_i$$

$$\hat{\Sigma}_{\text{RMCD}} = \frac{k_{\text{RMCD}}(m,n,v)}{m-1}\sum_{i=1}^{n} w_i (y_i - \hat{\mu}_{(\text{RMCD})})(y_i - \hat{\mu}_{(\text{RMCD})})'$$

- Again the scaling $k_{\text{RMCD}}(h,n,v)$ ensures consistency and unbiasedness

Robust RMCD Distances

- The outliers are revealed by their large (squared) Mahalanobis distances from the robust fit:

$$\hat{\mu}_{(\mathrm{RMCD})} = \frac{1}{h} \sum_{i \in y_{(\mathrm{RMCD})}} y_i$$

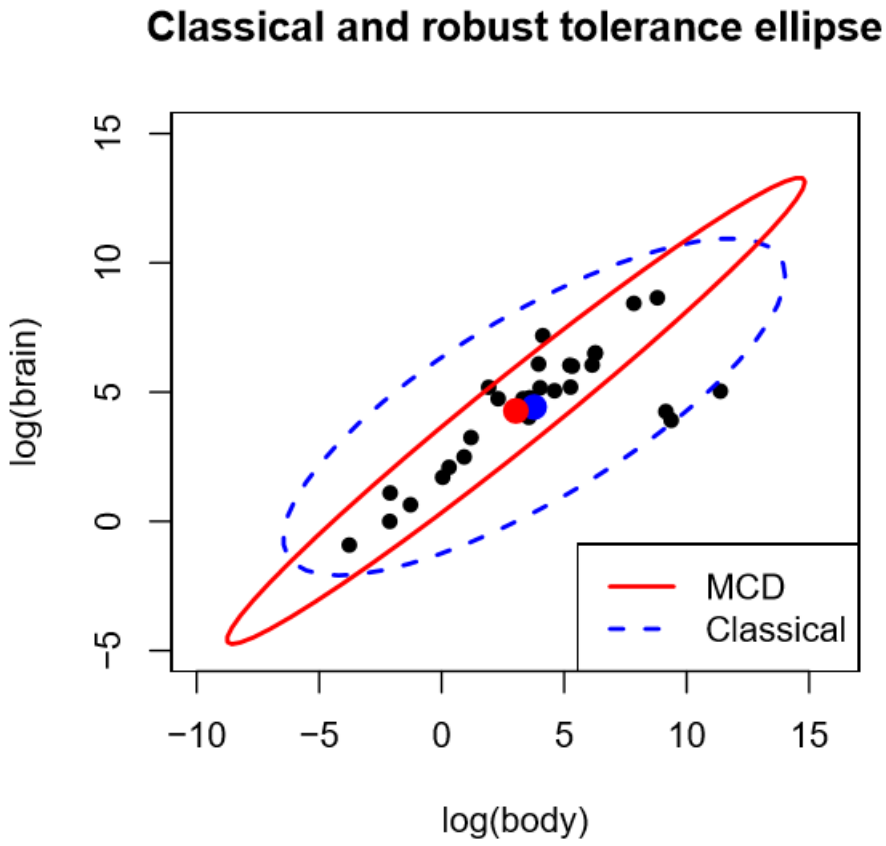
$$d^2_{i(\mathrm{RMCD})} = (y_i - \mu_{(\mathrm{RMCD})})' \hat{\Sigma}^{-1}_{(\mathrm{RMCD})} (y_i - \hat{\mu}_{(\mathrm{RMCD})})$$

- The robust distances **do not suffer from masking**
- The common suggestion is to use the 1% or 2.5% 5% cut-off values from the asymptotic Chi-squared distribution on v degrees of freedom. Ex. if 2.5% flag as outliers the obs. which do not belong to the robust tolerance ellipsoid

$$\{y; d^2_{i(RMCD)} \leq \chi^2_{v,0.975}\}$$

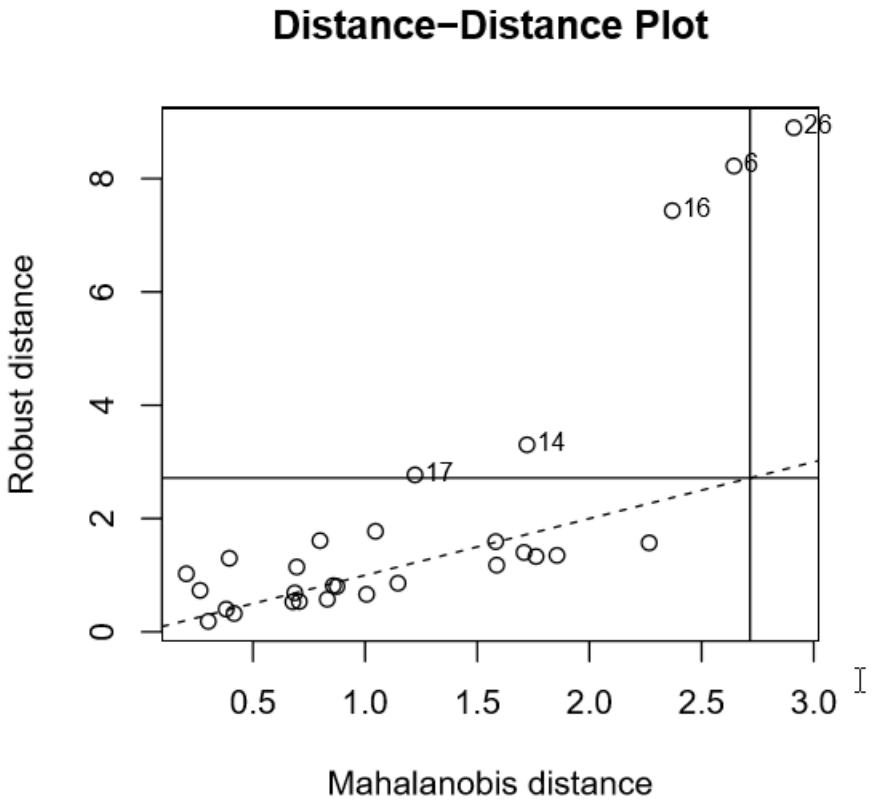
Outlier detection

- Outlier detection based on RMCD correctly flags the outliers in the animals data:



Distance-distance plot

- In dimensions p > 2, we cannot draw a scatterplot or a tolerance ellipsoid.
- To explore the differences between a classical and a robust analysis we can draw a distance-distance plot, which plots the points (MD*i*, RD*i*)



Computation of the MCD

- Exact algorithm: consider all h-subsets, compute the mean and covariance matrix of each, and retain the subset with smallest covariance determinant
- But: infeasible for large n or v
- Approximate algorithm: consider selected set of h-subsets. The most popular algorithm is FAST-MCD (Rousseeuw and Van Driessen, 1999). It uses random initial subsets.
- Recently a deterministic algorithm DetMCD has been developed, which is almost affine equivariant (Hubert et al., 2012).

Analysis of the test size of RMCD

- Monte Carlo estimate (50,000 simulations) of empirical size of the reweighted MCD outlier detection rule under no contamination and using a Bonferroni correction
- Nominal simultaneous size: 1%

	$n = 50$	$n = 75$	$n = 100$	$n = 200$	$n = 500$
$v = 6$	**0.549**	**0.246**	**0.138**	**0.035**	**0.018**
$v = 10$	**0.947**	**0.606**	**0.321**	**0.059**	**0.021**
$v = 12$	**0.995**	**0.803**	**0.483**	**0.077**	**0.024**

- **Some correction is needed to reduce the number of false outliers in finite samples!**

Approach 1

- Calibrate the **cut-off values** of the distribution of robust distances, not just its first two moments (Cerioli, Riani, and Atkinson, Stat. & Comp. 2009)
- Use Beta and F distribution (Cerioli, JASA 2010)

Performance of corrected MCD

- **nominal size of 1%** for the test of no outliers
- Max breakdown: $h = \lfloor (n + v + 1)/2 \rfloor$
- Trimming at 0.975 in the reweighting step
- 5000 simulations for each combination of n and v.

	$n = 40$	$n = 60$	$n = 90$	$n = 125$	$n = 200$	$n = 400$
$v = 5$	0.017	0.017	0.015	0.013	0.011	0.010
$v = 10$	0.054	0.025	0.014	0.012	0.012	0.008
$v = 15$	0.084	0.030	0.013	0.014	0.013	0.010

FS: flexible power improvement

- The Forward Search (FS) relies on a fully-iterative **adaptive trimming** scheme:
- **Order the data** by closeness to the assumed model (for outlier detection: $N(\mu, \Sigma)$)
- **Start** with a small subset of m_0 observations
- **Move Forward:** increase the number of observations m used for fitting the model.
- The choice of the new subset (of cardinality $m + 1$) is based on the **distances** computed at step m
- Continue until $m = n$
- Outliers and other observations not following the general structure **enter at the end** and can be clearly identified

The FS details at step m

- $S(m)$:= fitting subset of m observations at **step m of the FS**
- We compute the estimates of the centroid and covariance matrix from S(m), $\hat{\mu}(m)$ and $\hat{\Sigma}(m)$
- These estimates yield n **squared distances**

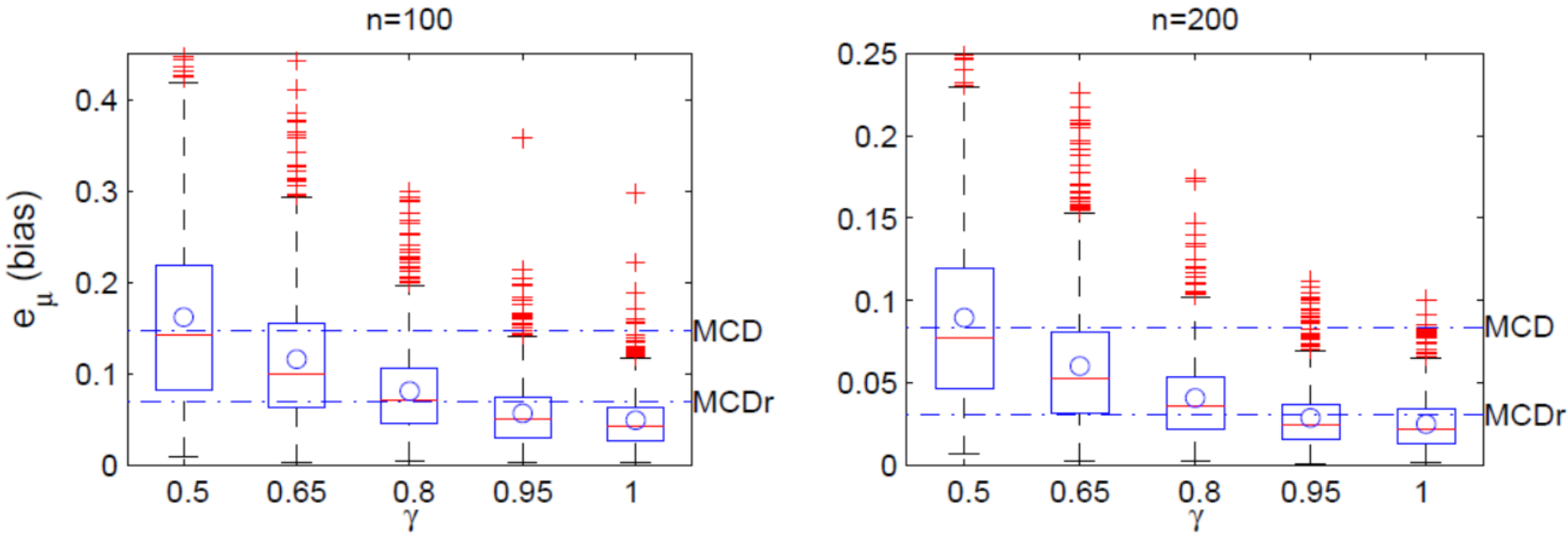
$$d_i^2(m) = \{y_i - \hat{\mu}(m)\}' \hat{\Sigma}(m)^{-1} \{y_i - \hat{\mu}(m)\} \qquad i = 1, \ldots, n$$

- Order these squared distances and take the observations corresponding to the $m + 1$ smallest as the new subset $S(m + 1)$

Theoretical results

- Cerioli, Farcomeni, and Riani, JMVA (2014) show that $\hat{\mu}(m)$ and $\hat{\Sigma}(m)$ are **strongly consistent** under the null model and have **breakdown point** $1 - m/n$ under contamination:
- **The FS yields consistent high-breakdown estimators, but with adaptive breakdown point**

Empirical Performance of FS Estimators Comparison with MCD and RMCD



- Boxplots of the values of the **squared bias** for the FS estimator of location, as a function of $\gamma = m/n$, for $n = 100$ (left) and $n = 200$ (right). The circles over the boxplots denote the average values
- The horizontal dashed-dotted lines are associated with the squared bias for the MCD location estimator (upper line) and the Reweighted MCD (MCDr) location estimator (lower line)

The FS for outlier detection

- **Importance of monitoring : a wealth of diagnostics can be computed and displayed along the search.**
- The main tool for **outlier detection** is the forward plot of the **Minimum distance among units outside the subset (min MD)**

$$d_{\min}(m) = \min d_i(m) \qquad i \notin S(m)$$

- If observation $[m + 1]$ is an outlier, its distance will be large compared to the maximum distance of the m observations in $S(m)$: **peak in the forward plot** of $\mathrm{d}_{\min}(m)$.

Scaled MD

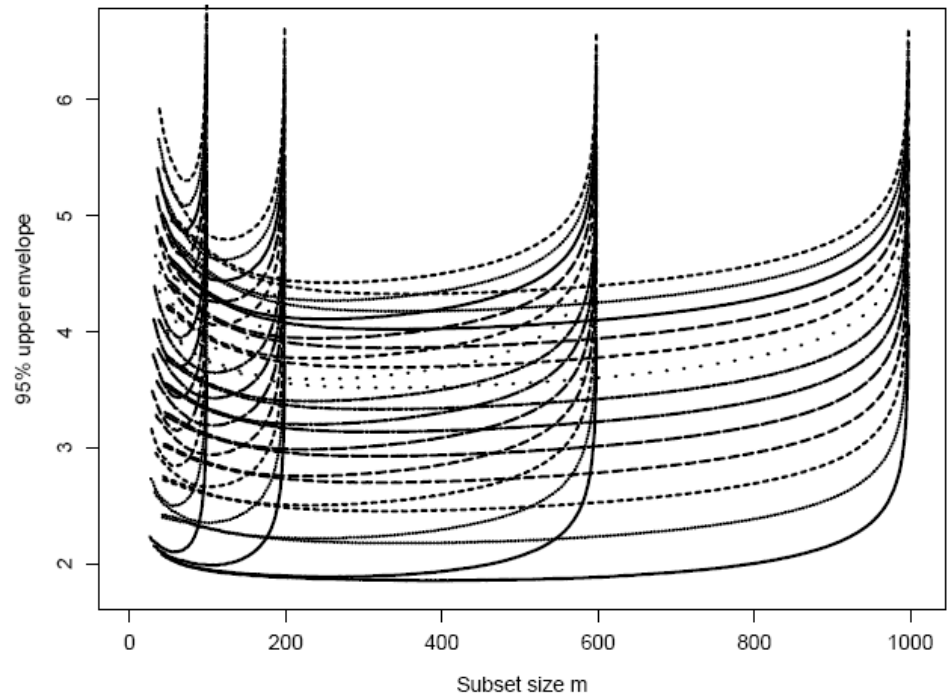
- In regression we work with scaled residuals

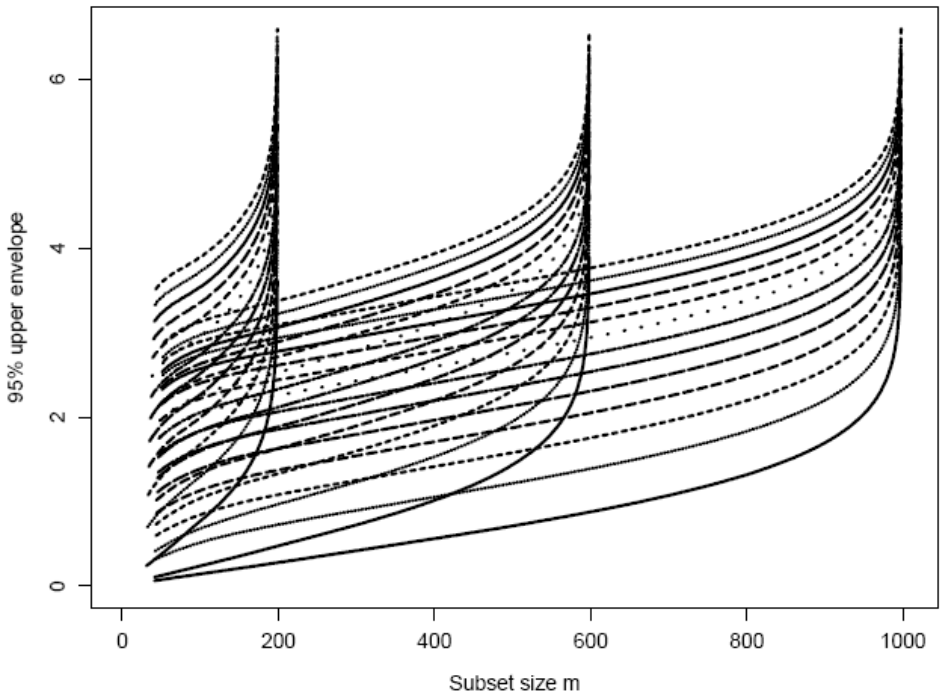
- In multivariate analysis we scale MD as

$$d_i^{\mathrm{SC}}(m) = d_i(m) \times \left(|\hat{\Sigma}(m)| / |\hat{\Sigma}(n)| \right)^{1/2v}$$

Dependence on *v* and *n*

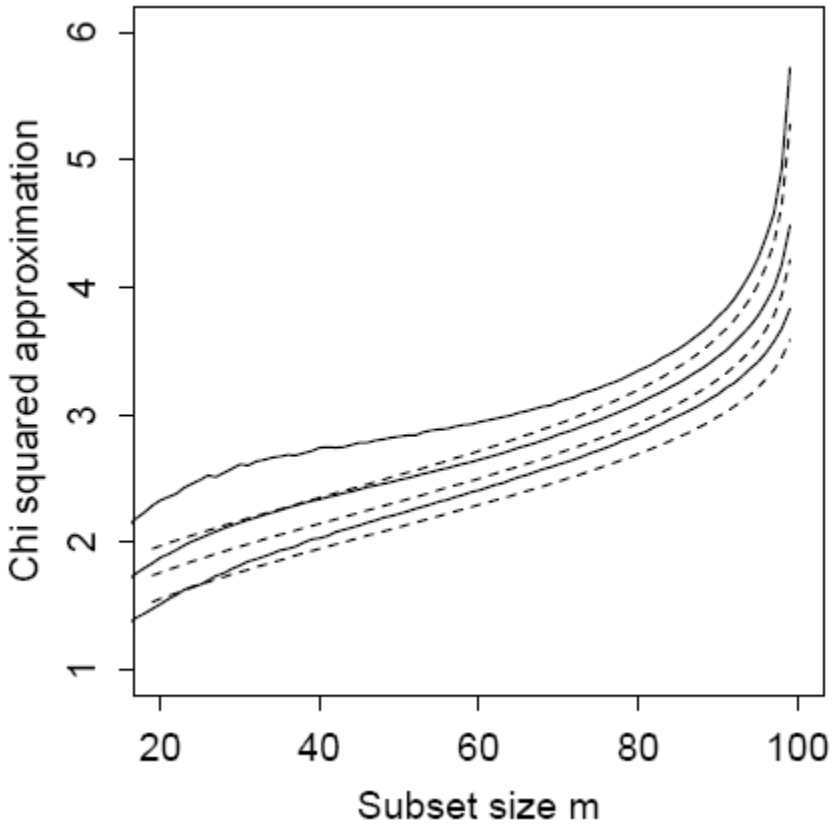
Unscaled distances Scaled distances

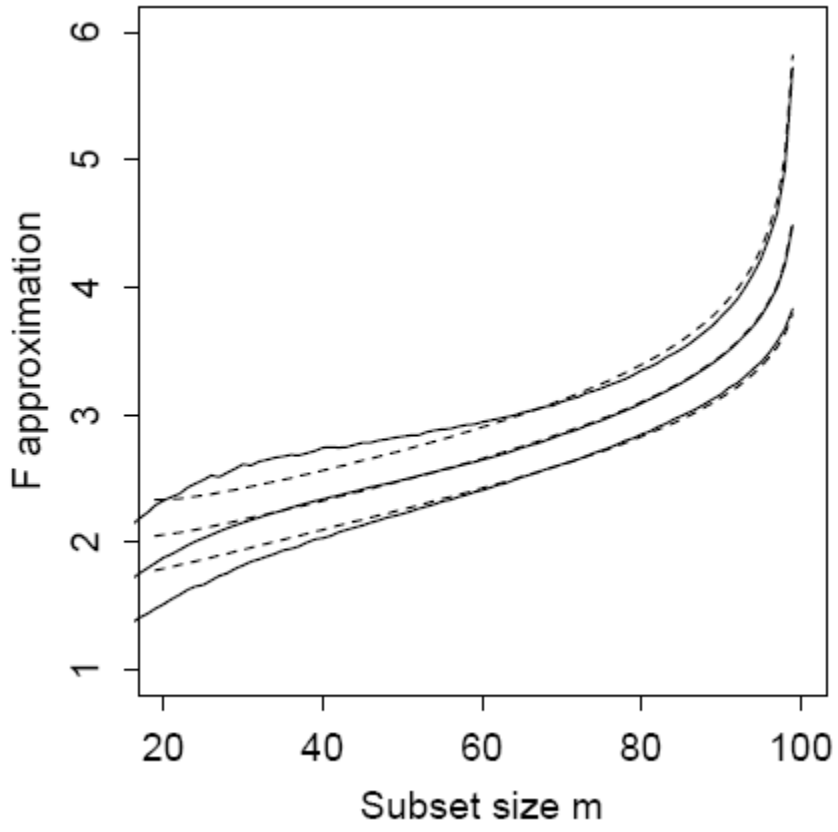




- 95% points of the empirical distribution of the min. MD for sample sizes n=100, 200, 600 and 1000; v=1, 2, …13

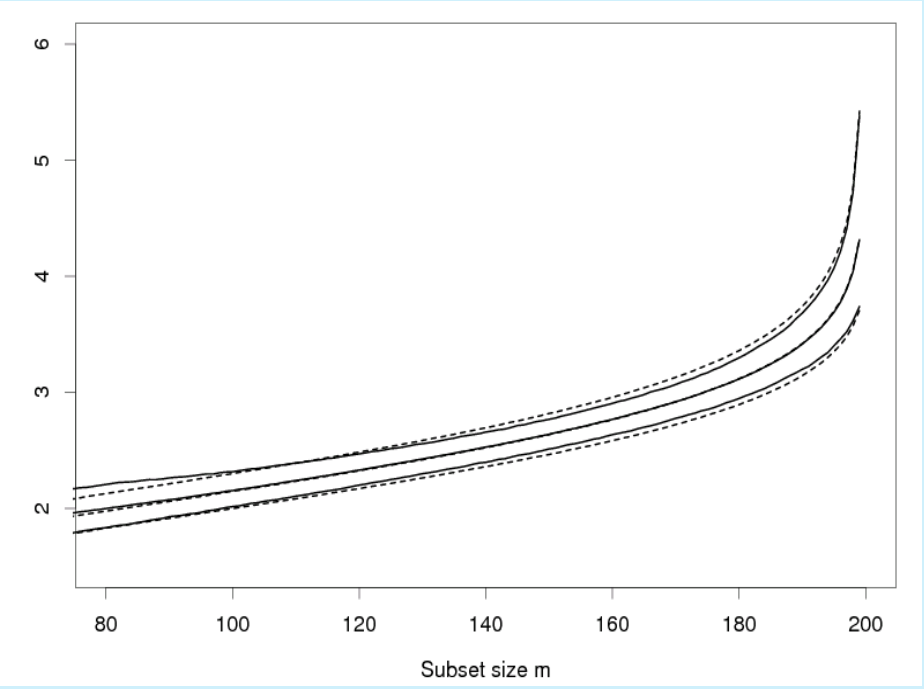
Approximation based on order statistics



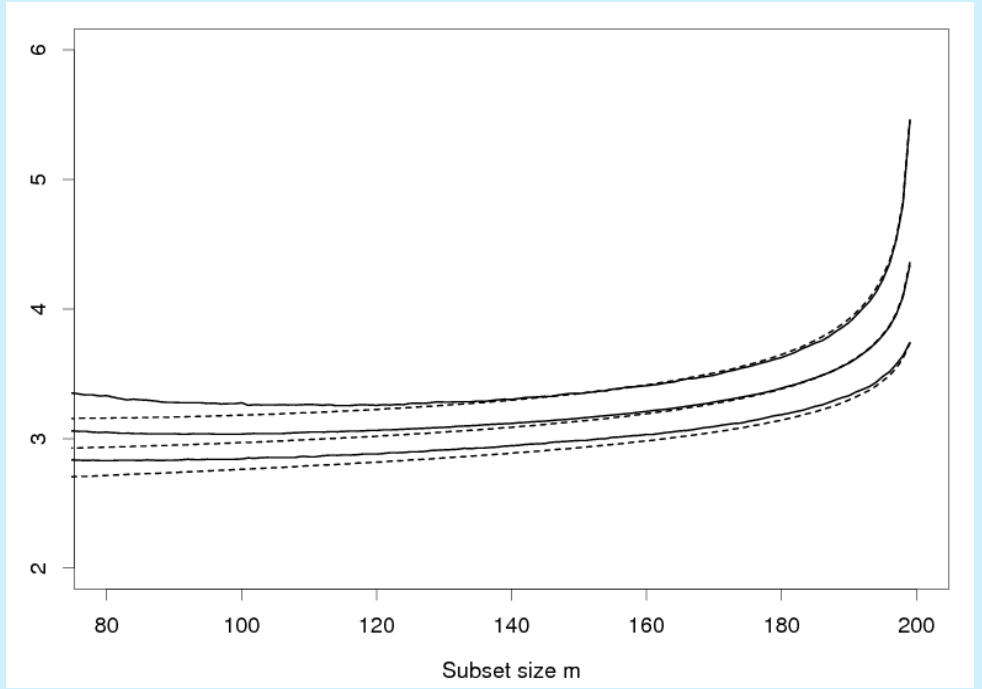


Comparison of 1%, 50% and 99% asymptotic envelopes for scaled distances. Continuous lines: envelopes found by simulation (n=100 and v=6)

Approximation based on order statistics n=200 v=5

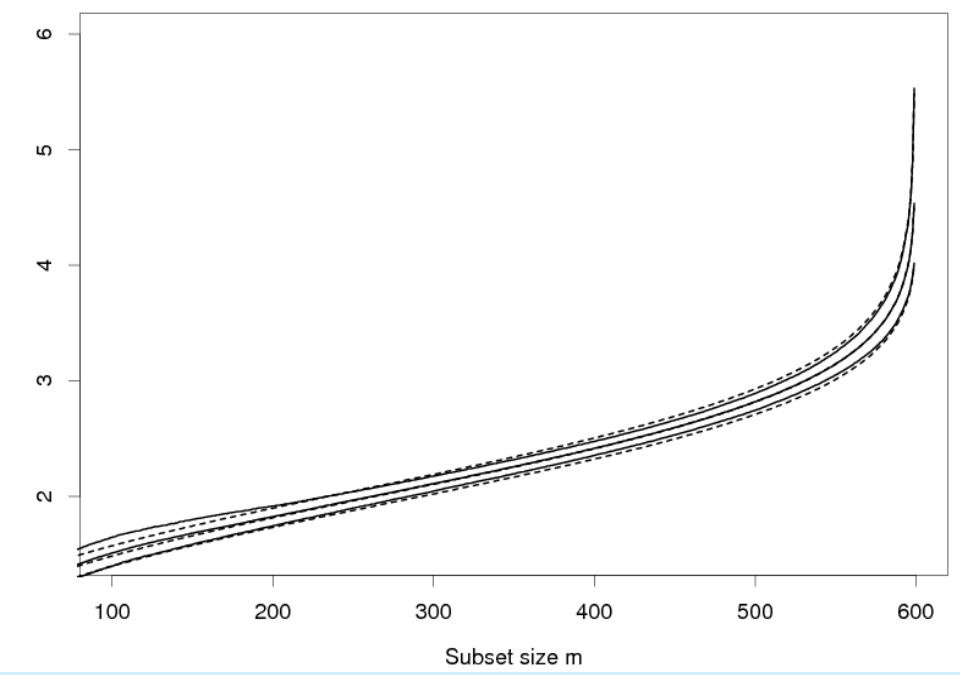


Scaled distances

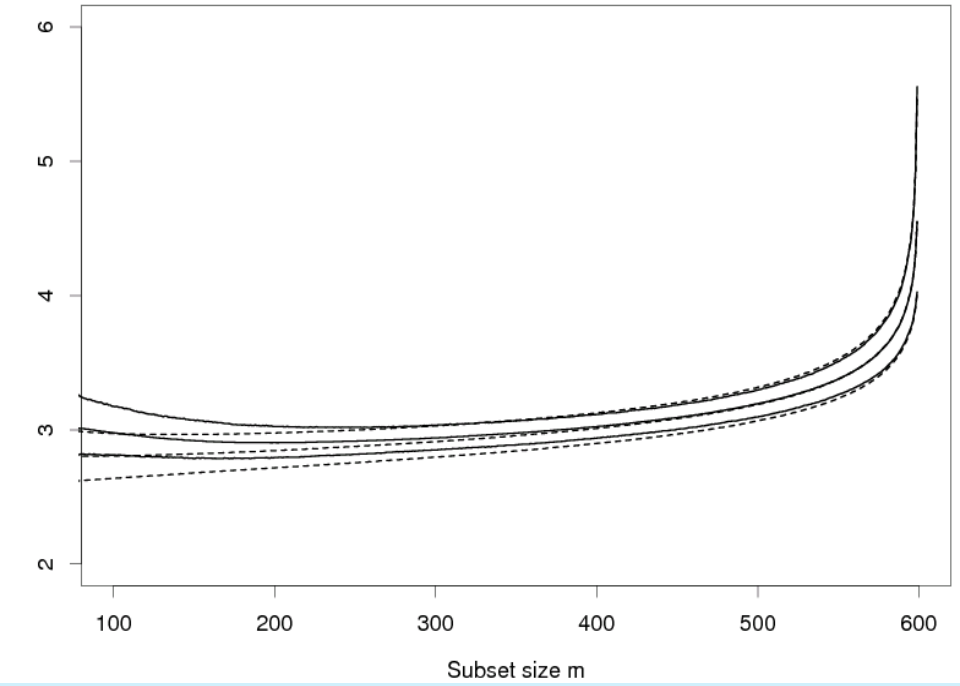


Unscaled distances

Approximation based on order statistics n=600 v=5



Scaled distances



Unscaled distances

Swiss heads data

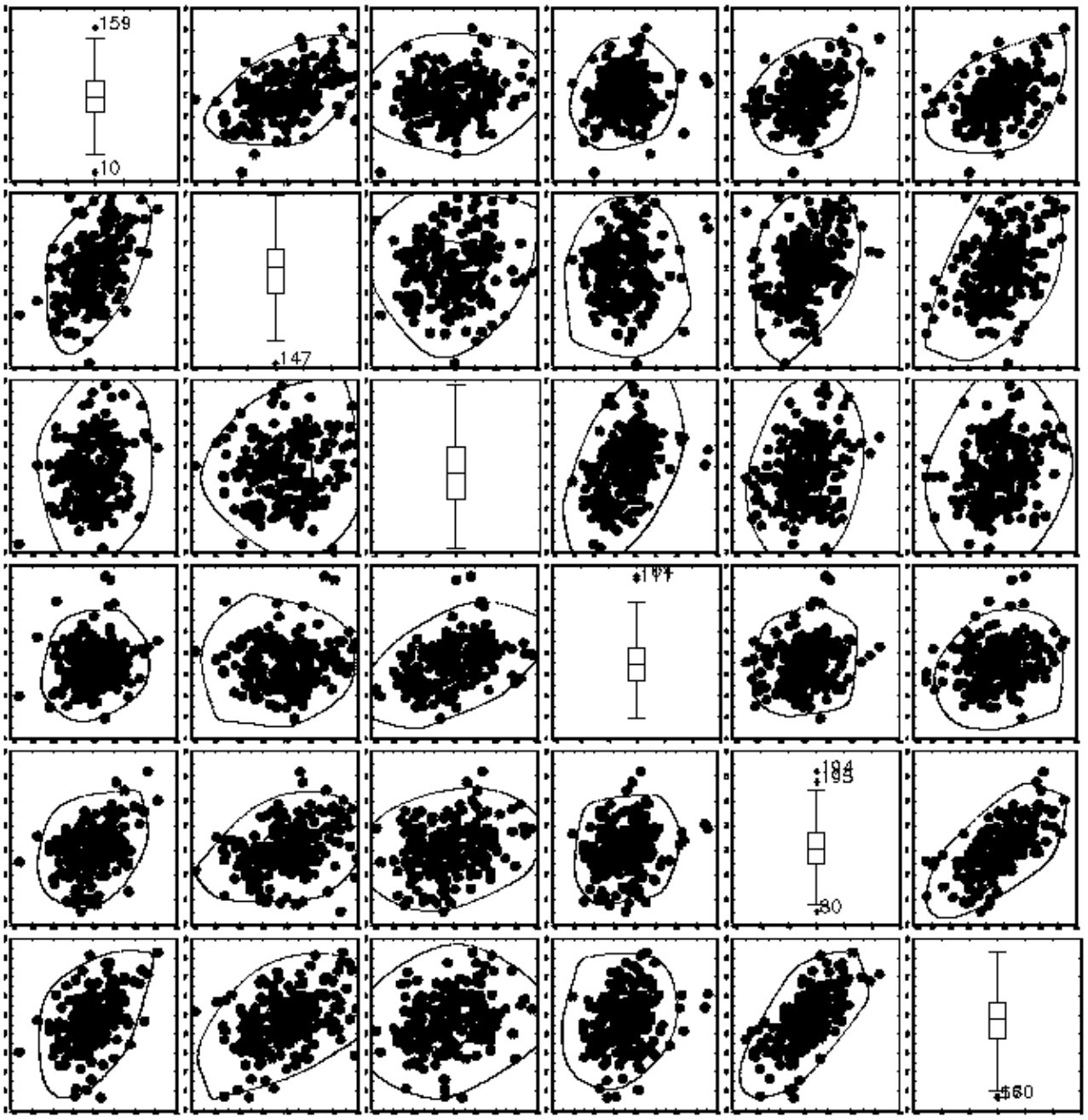
- Six readings on the dimensions of the heads of 200 twenty years old soldiers

y_1: minimal frontal breadth
y_2: breadth of angulus mandibulae
y_3: true facial height
y_4: length from glabella to apex nasi
y_5: length from tragion to nasion
y_6: length from tragion to gnathion.

Swiss heads data

- Final purpose: to study the variability in size and shape of young men in order to help to design a new protection mask
- Choice of the initial subset: we find an initial subset of $m0$ observations from the intersection of units inside a robust bivariate contour for each pair of variables

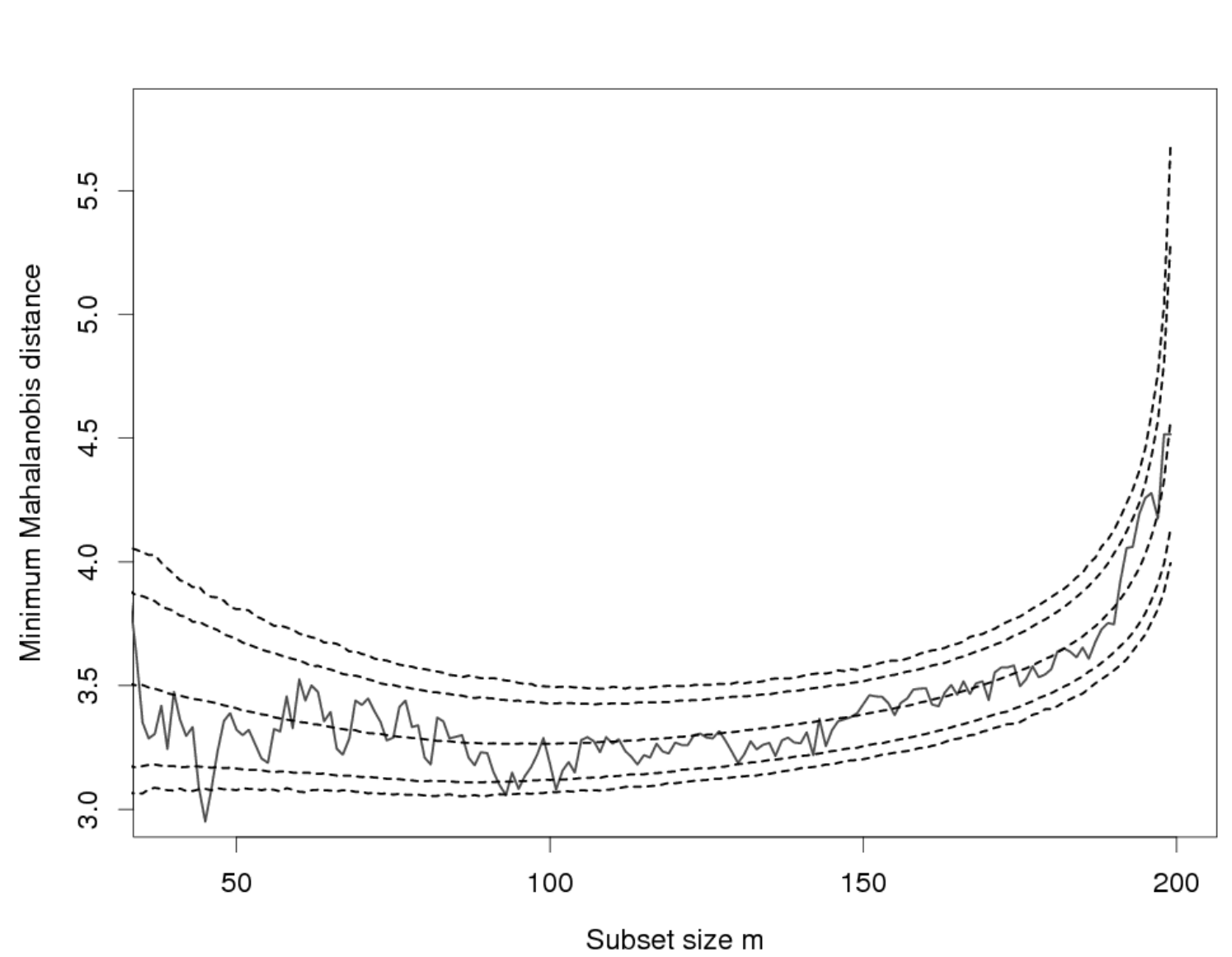
Swiss heads data: SPM with univariate and bivariate boxplots



Remark on the initial definition of starting point

- Does not involve complicated iterative procedures
- The size of the initial subset can easily be decreased or increased by changing the value of the outer contour
- We can easily try several starting points and check whether the final part of the search is the same
- We can force the starting point

Monitoring min MD with envelopes



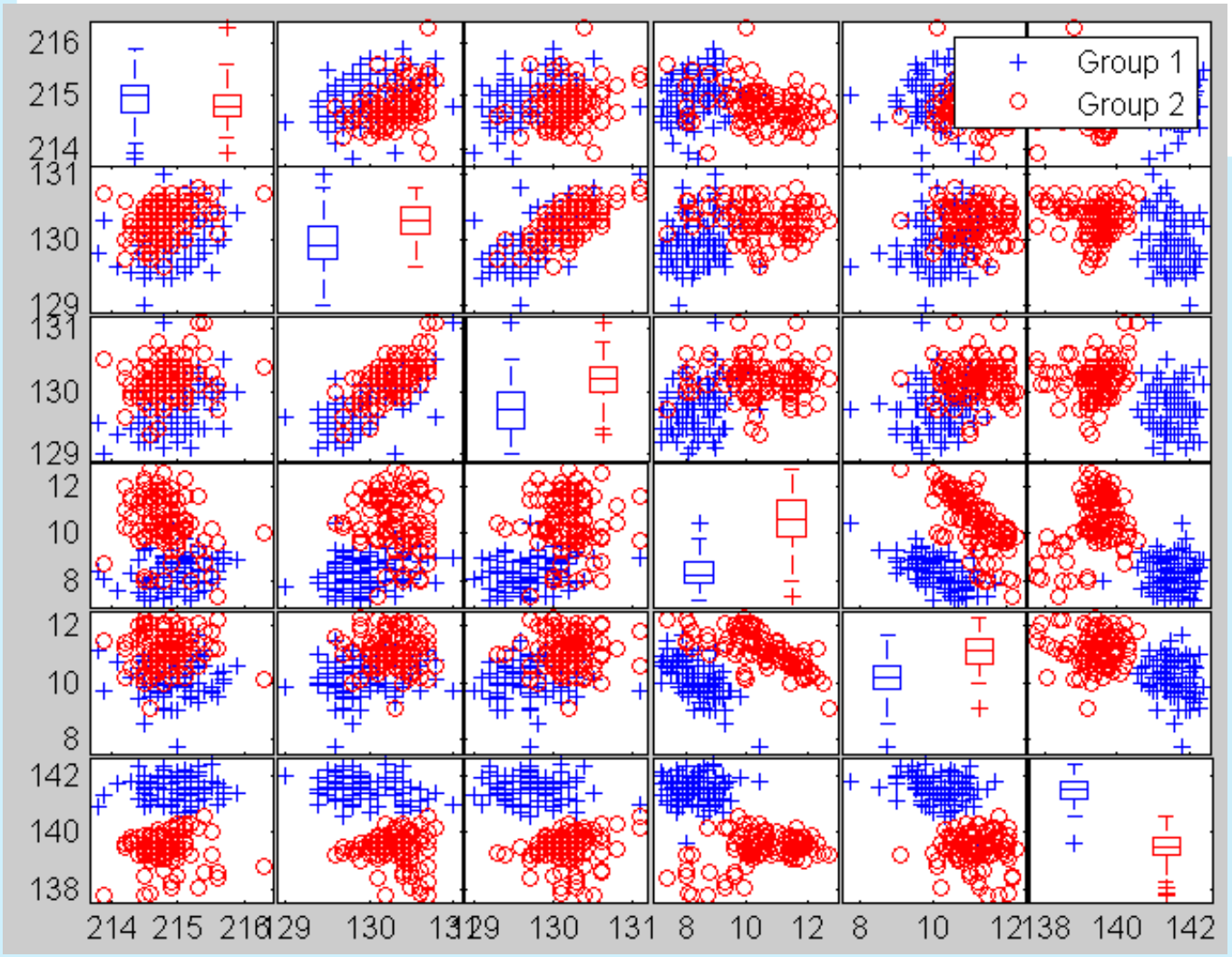
Swiss bank notes

- 6 variables are measurements of the size of the bank notes
- 100 of which are genuine and 100 are forged
- Some complications:
  - Some of the notes in either group may have been misclassified
  - Forged notes may not form a homogeneous group

y_1: length of bank note near the top
y_2: left-hand height of bank note
y_3: right-hand height of bank note
y_4: distance from bottom of bank note to beginning of patterned border
y_5: distance from top of bank note to beginning of patterned border
y_6: diagonal distance

Swiss Banknotes

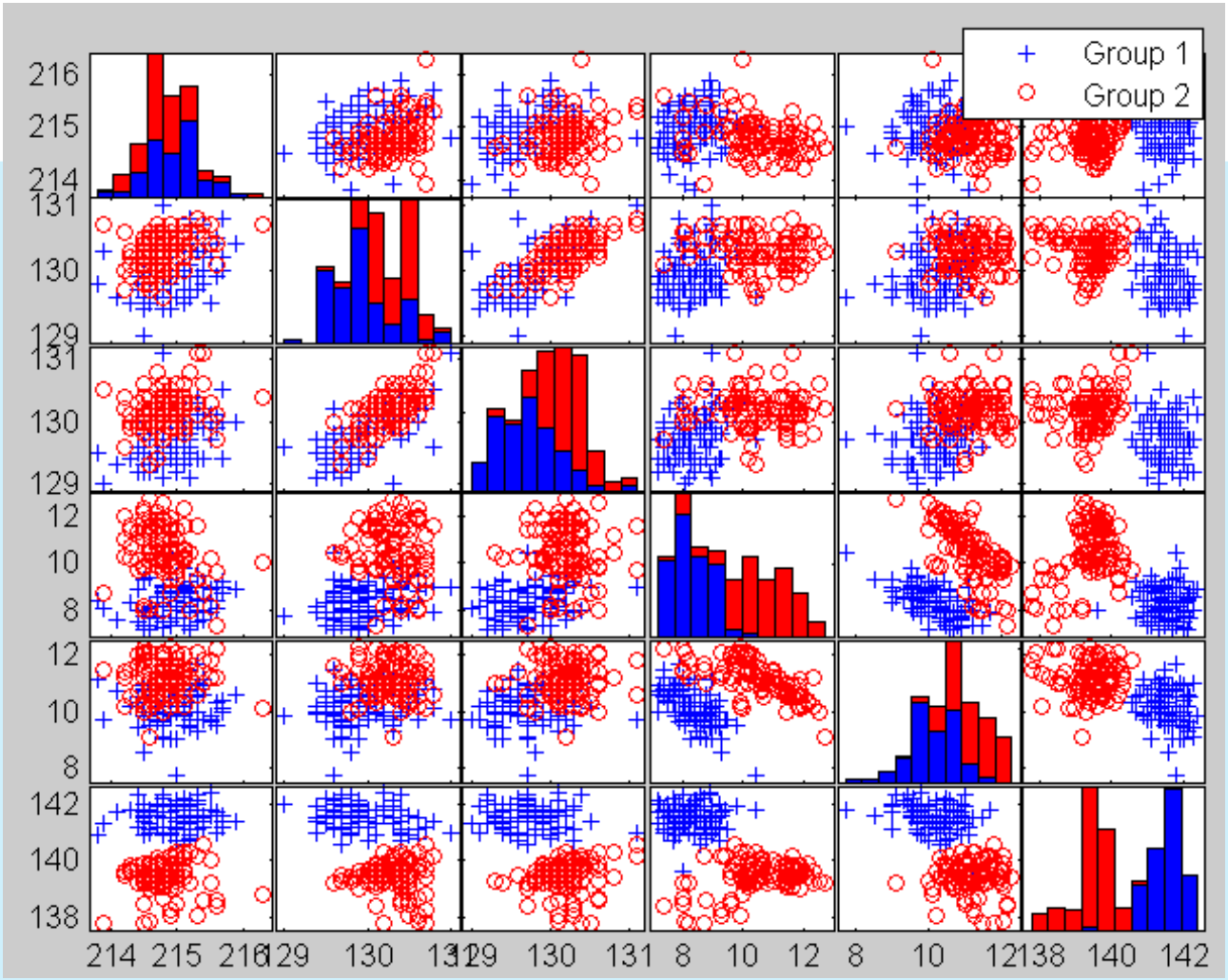
- 100 are genuine and 100 are forged (n=200 v=6)
- Two(?) populations (genuine and forged notes), but with **several outliers** (different forgers?)



Both extreme and intermediate outliers

Swiss Banknotes

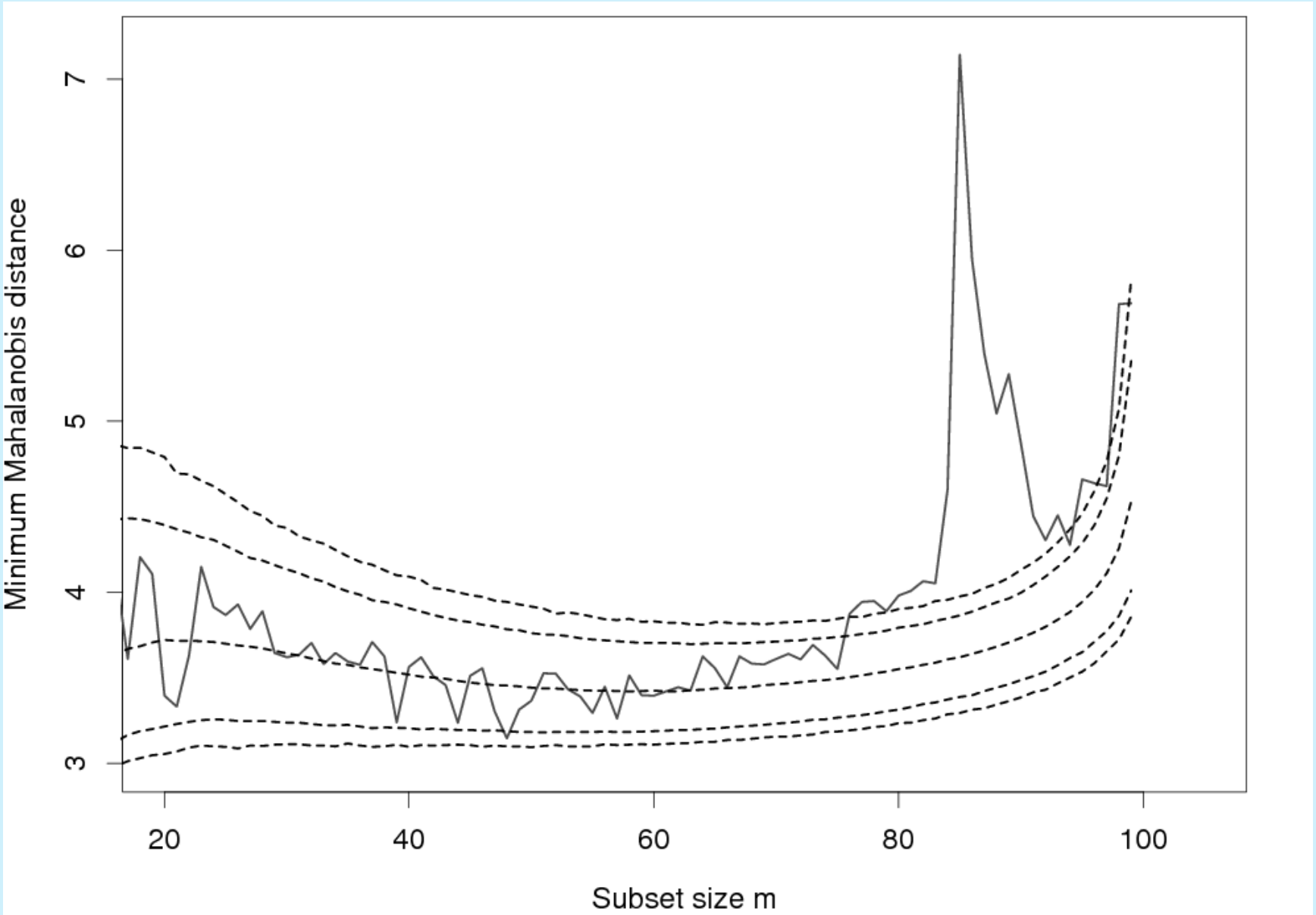
- 100 are genuine and 100 are forged
- Two(?) populations (genuine and forged notes), but with **several outliers** (different forgers?)



Both extreme and intermediate outliers

Analysis of the group of 100 fake banknotes

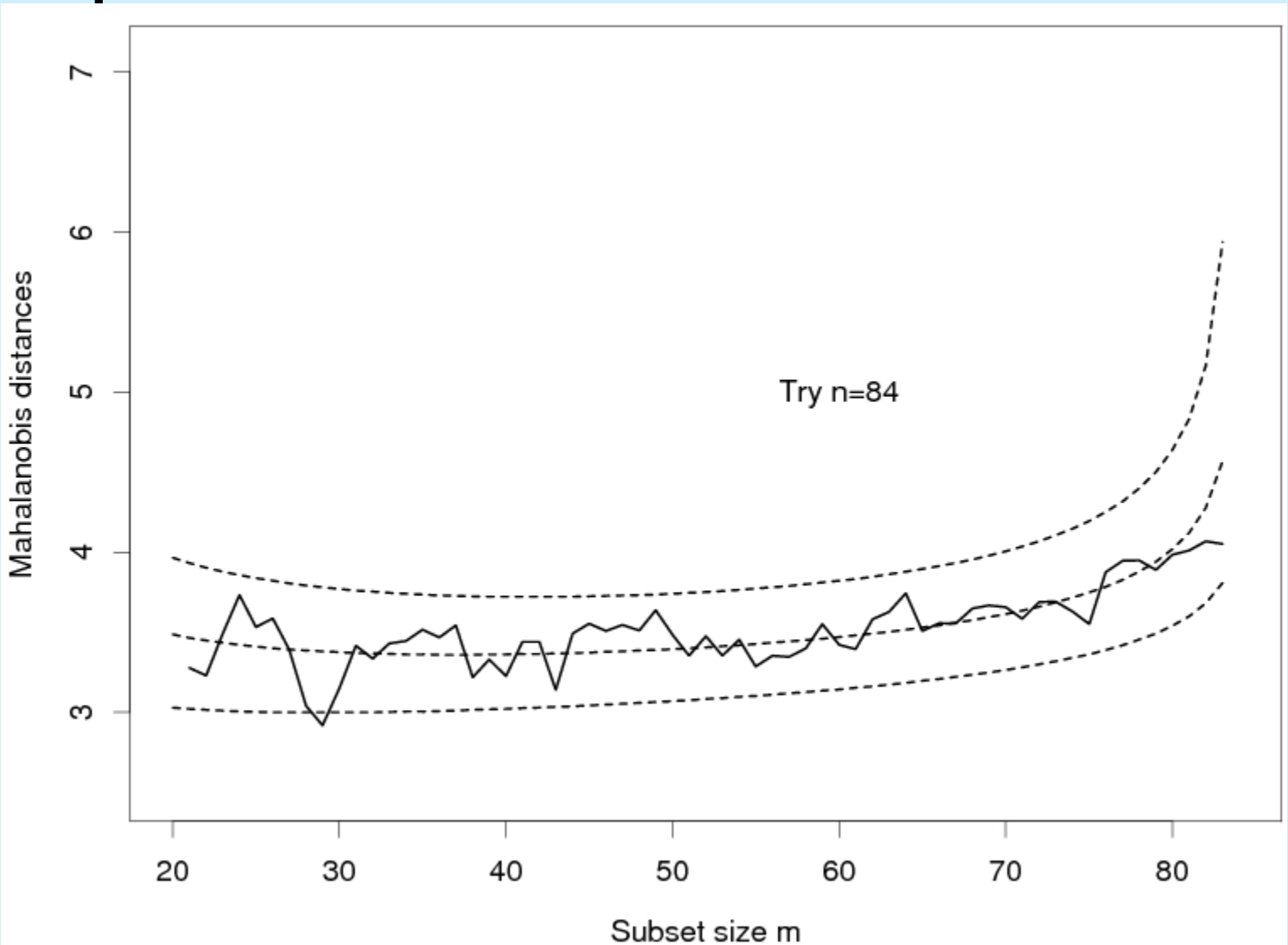
Monitoring min MD with envelopes



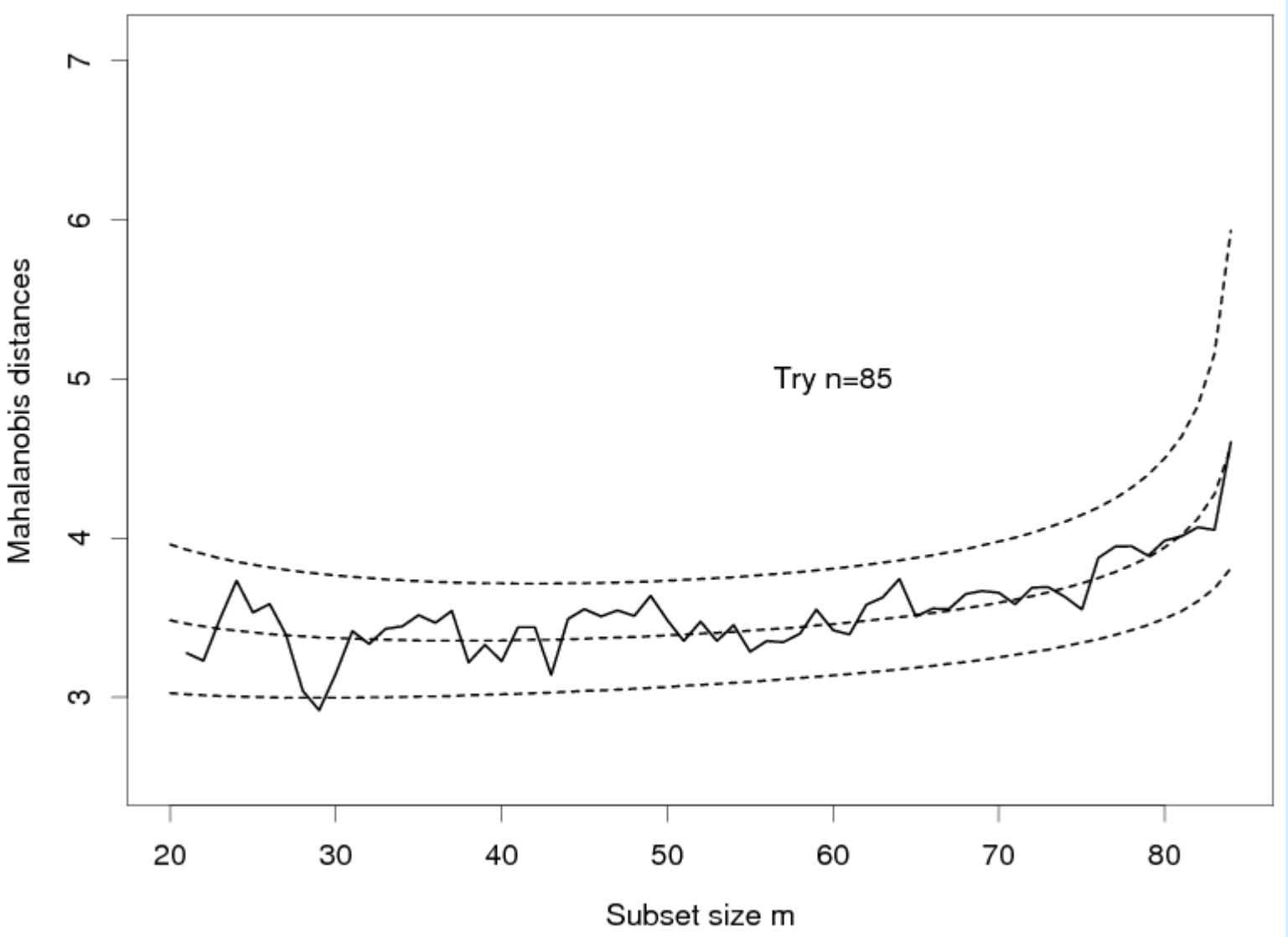
Strategy for outlier detection

- Once a signal takes place ($m=m^*$) start superimposing 99% envelopes using $n=m^*-1$, m^*, m^*+1 up to when the trajectory is inside the threshold

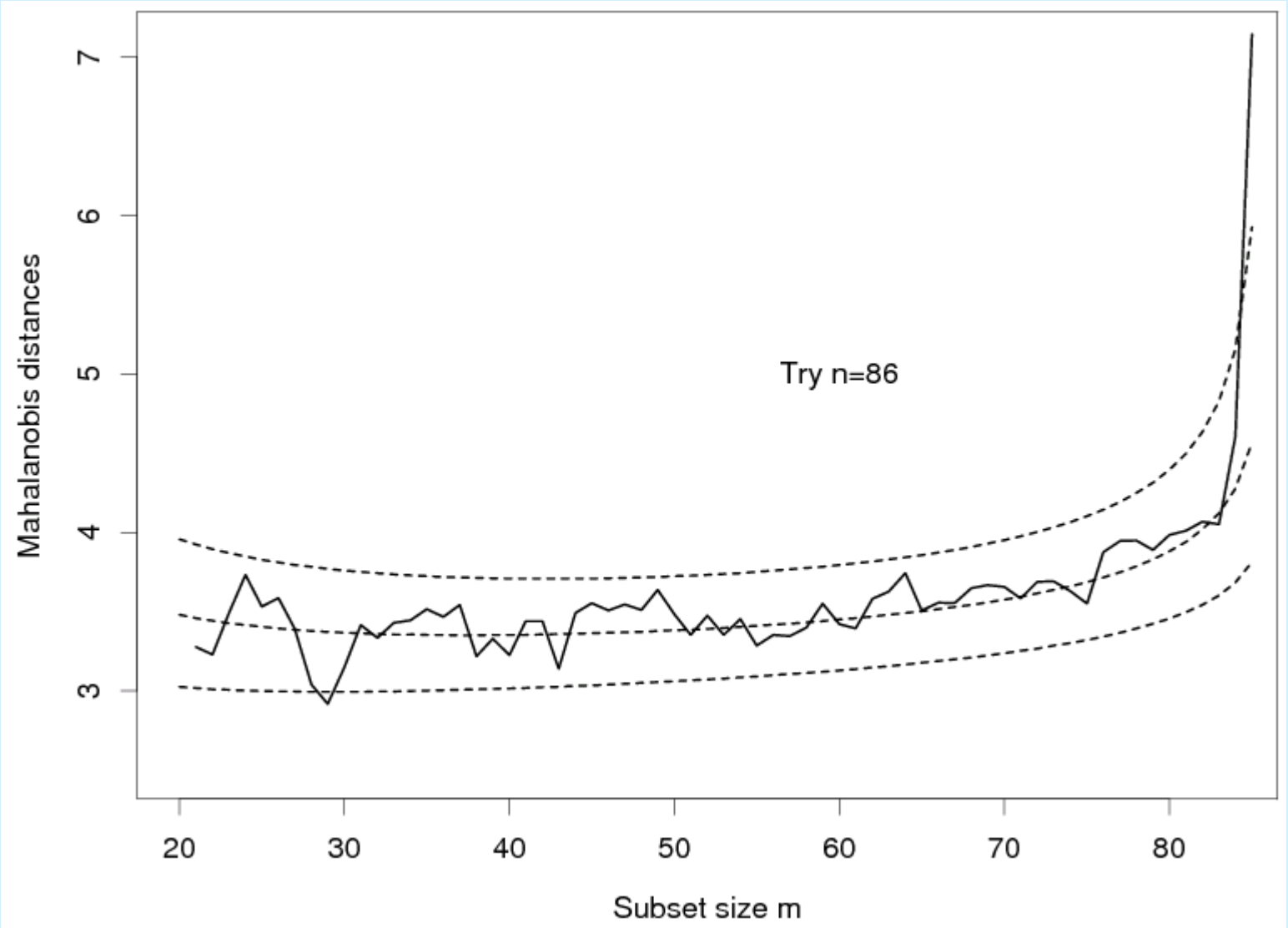
SBN: superimposed envelopes at step *n*=84



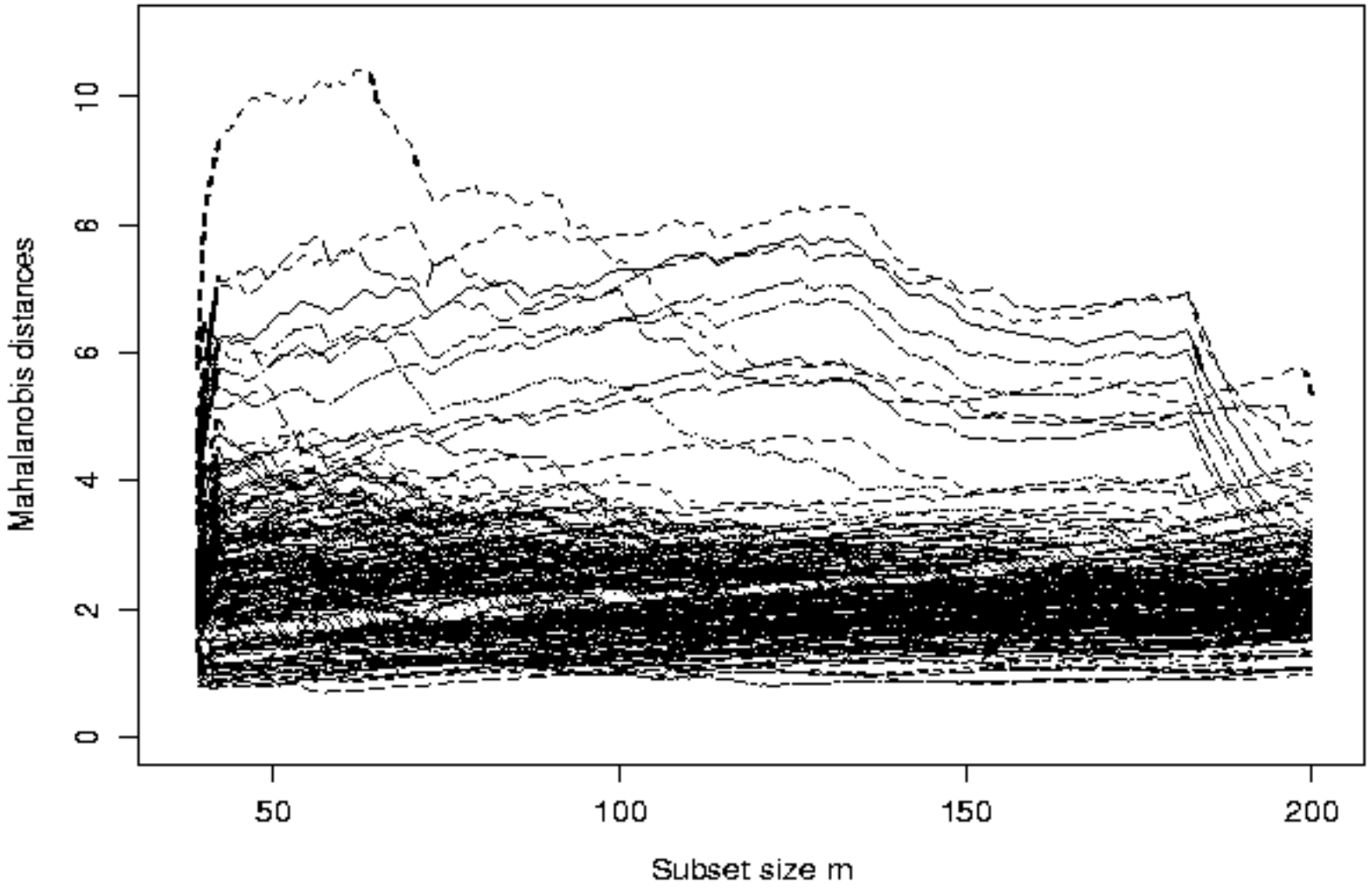
SBN: resuperimposed envelopes at step *n*=85



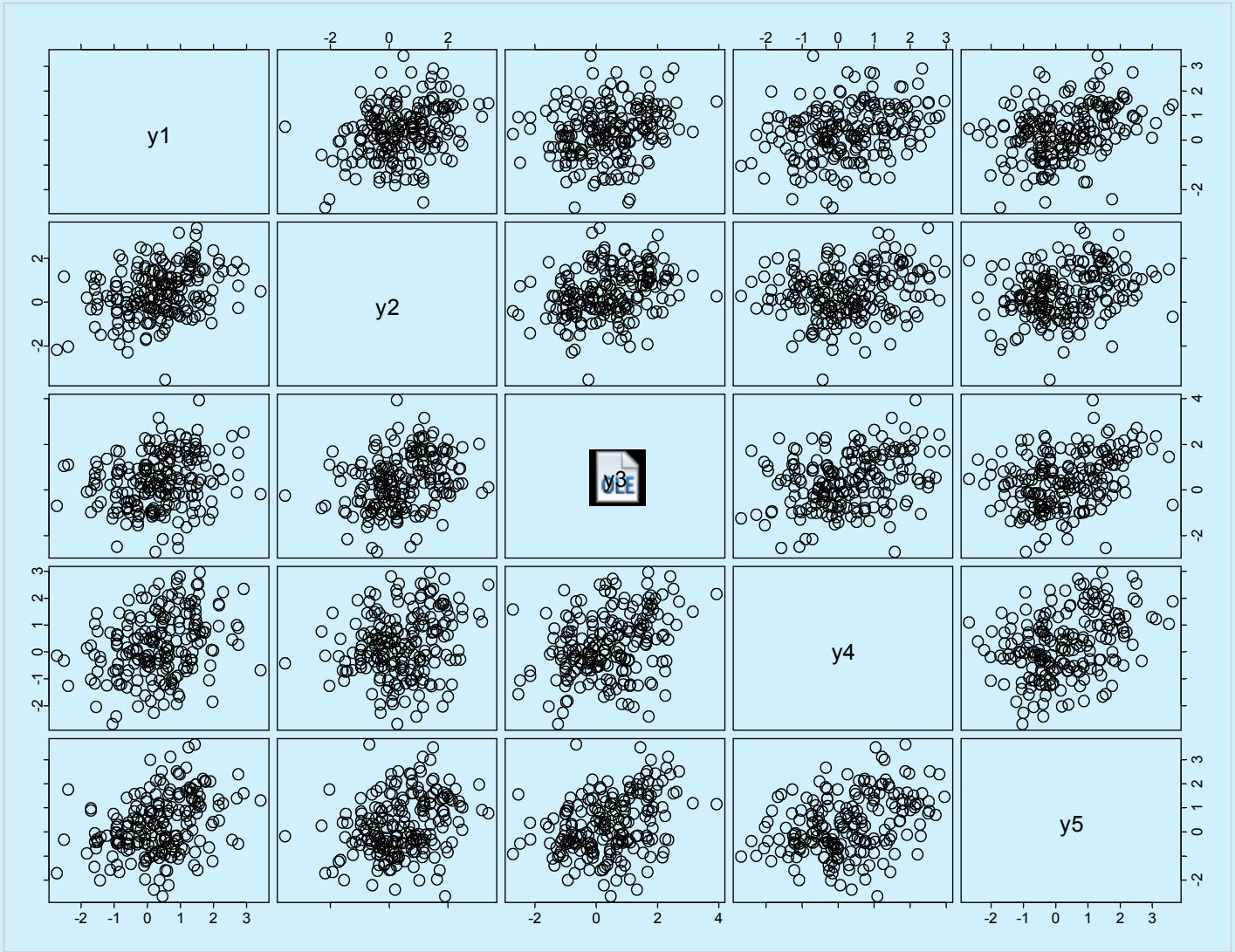
SBN: resuperimposed envelopes at step *n*=86



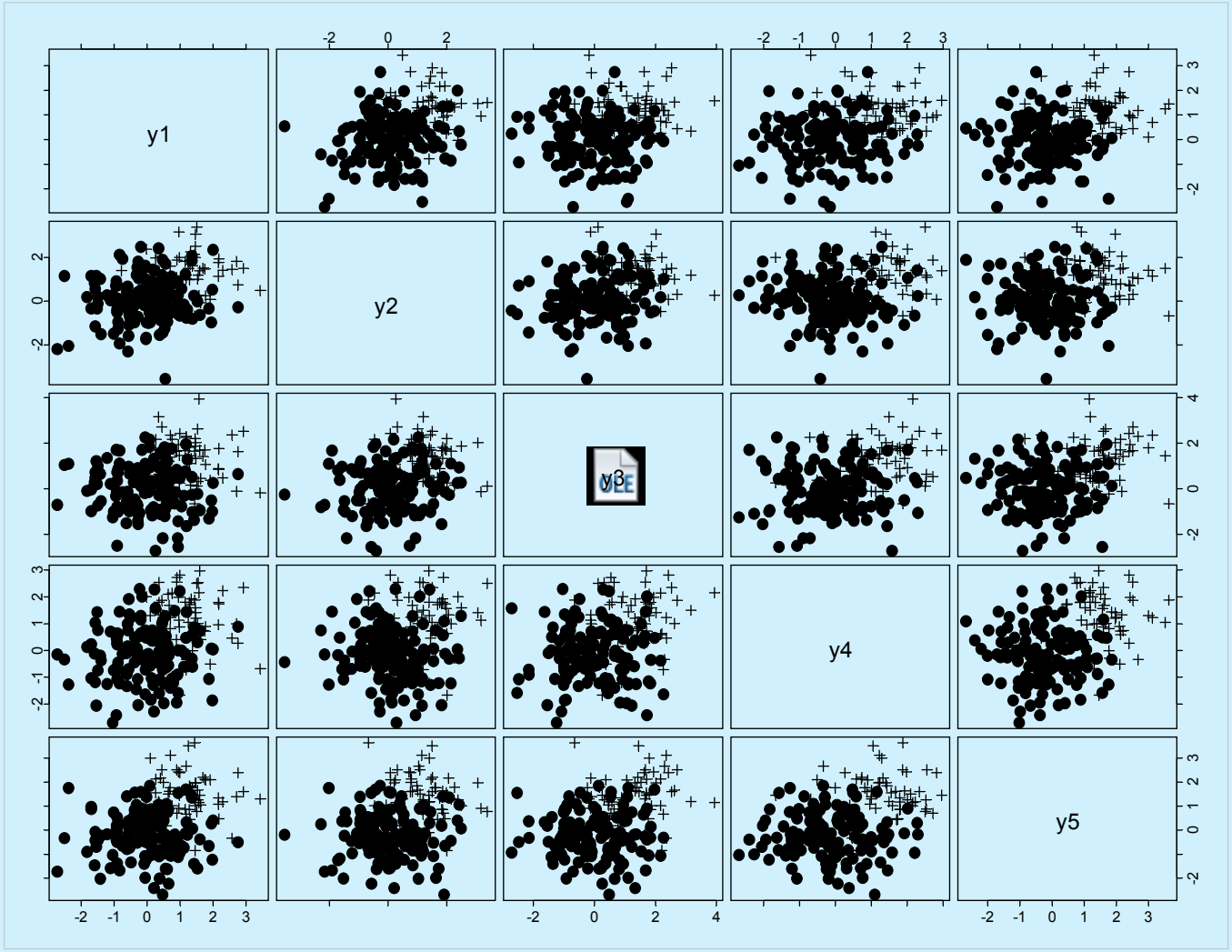
Forward plot of scaled MD



Example with simulated data: n=200 p=5, 30% contamination (first 60 obs.) Level shift=1.2

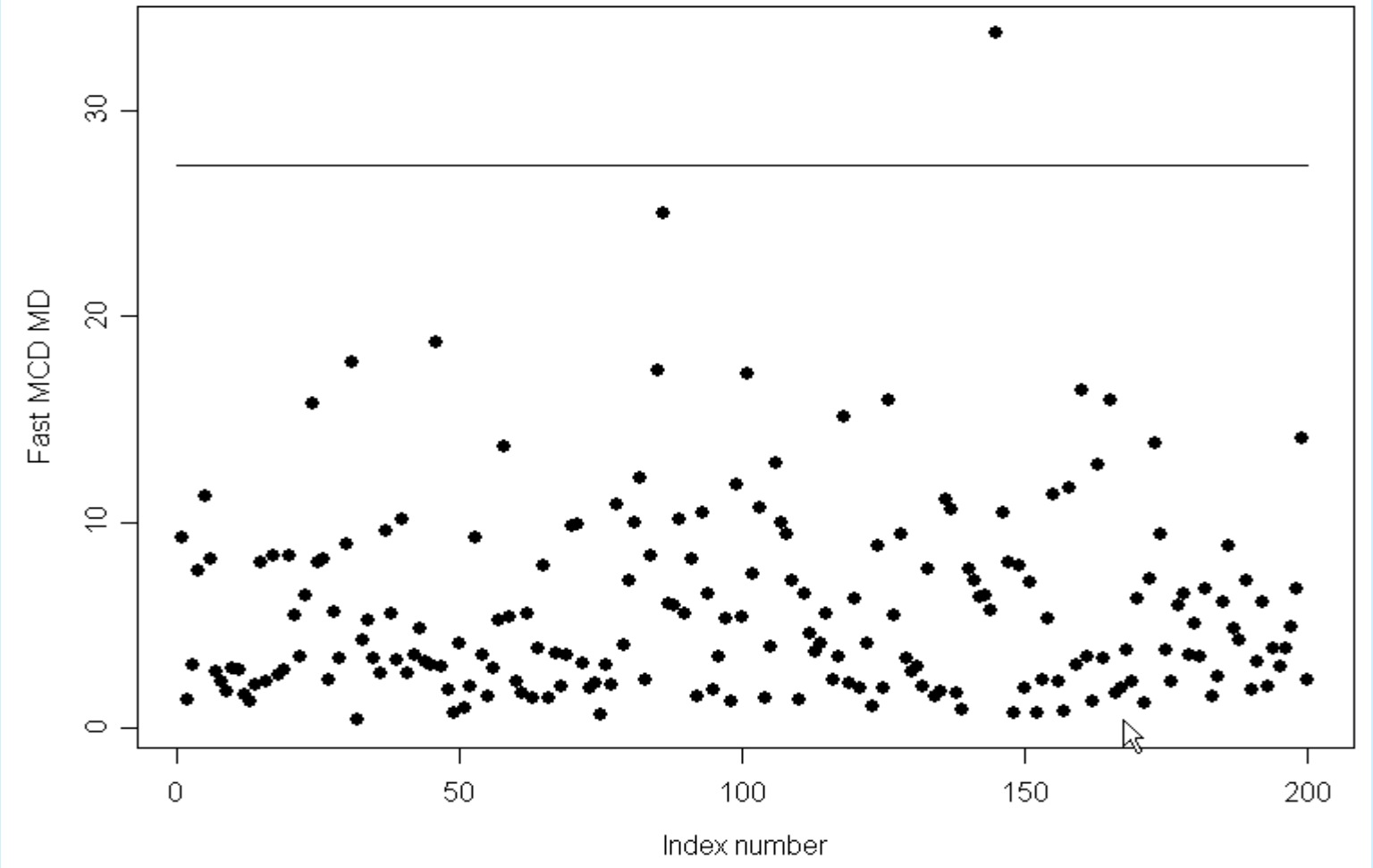


Example with simulated data: n=200 p=5, 30% contamination (first 60 obs.) Level shift=1.2



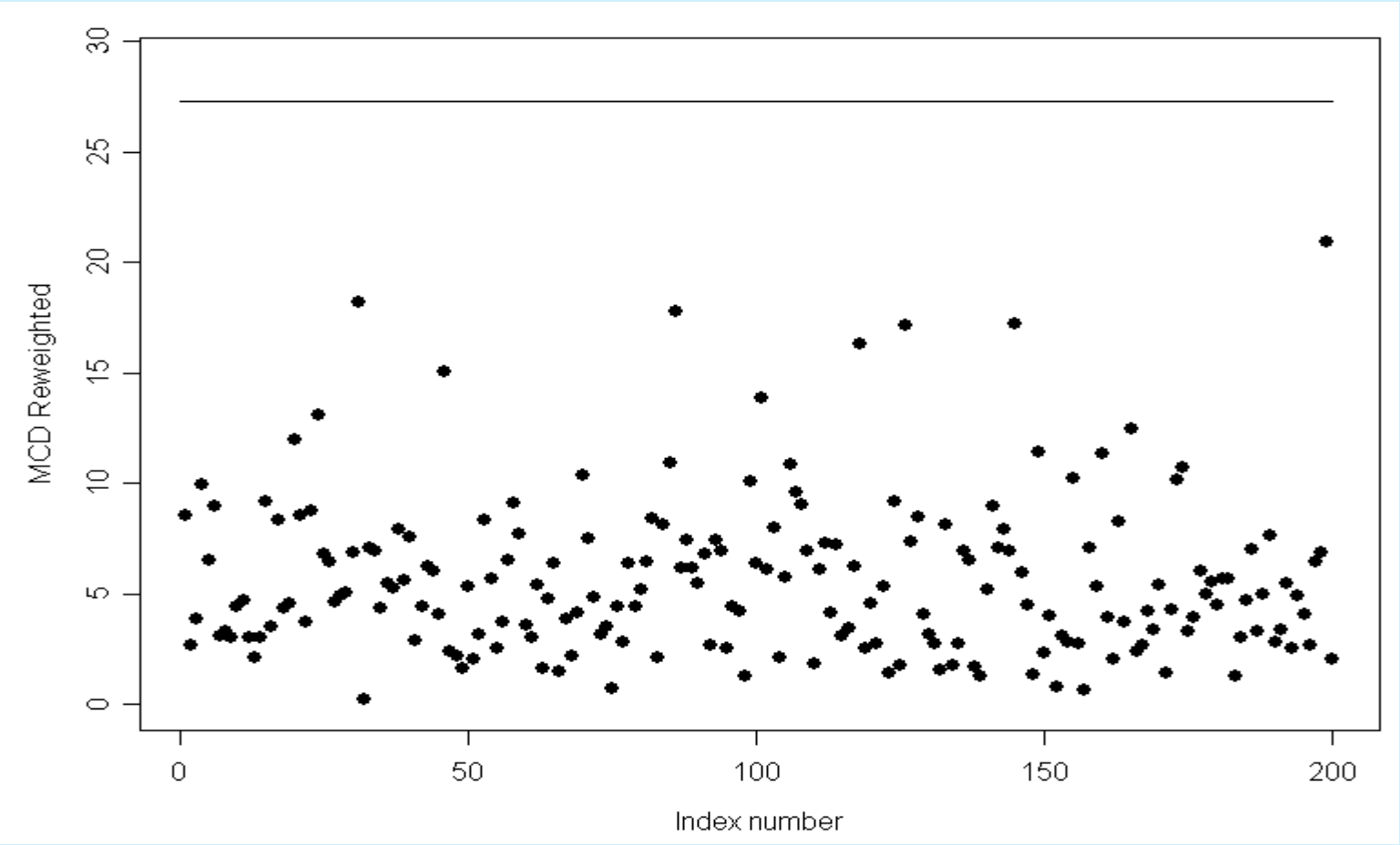
n=200 *p*=5 30% contamination Level shift=1.2

Output from FAST MCD (consistency correction + Pison correction - Real α=0.3291)

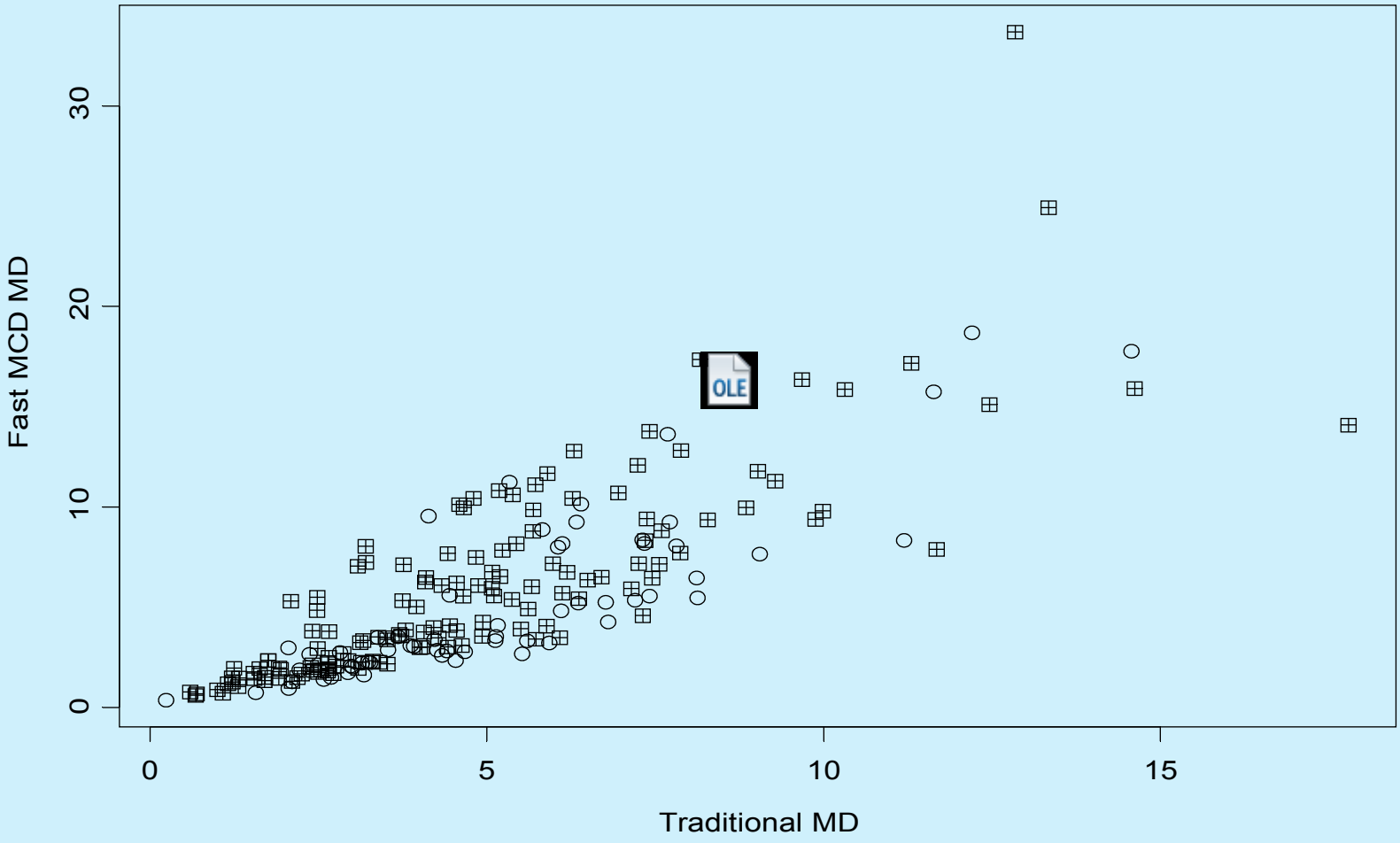


n=200 *p*=5 30% contamination Level shift=1.2

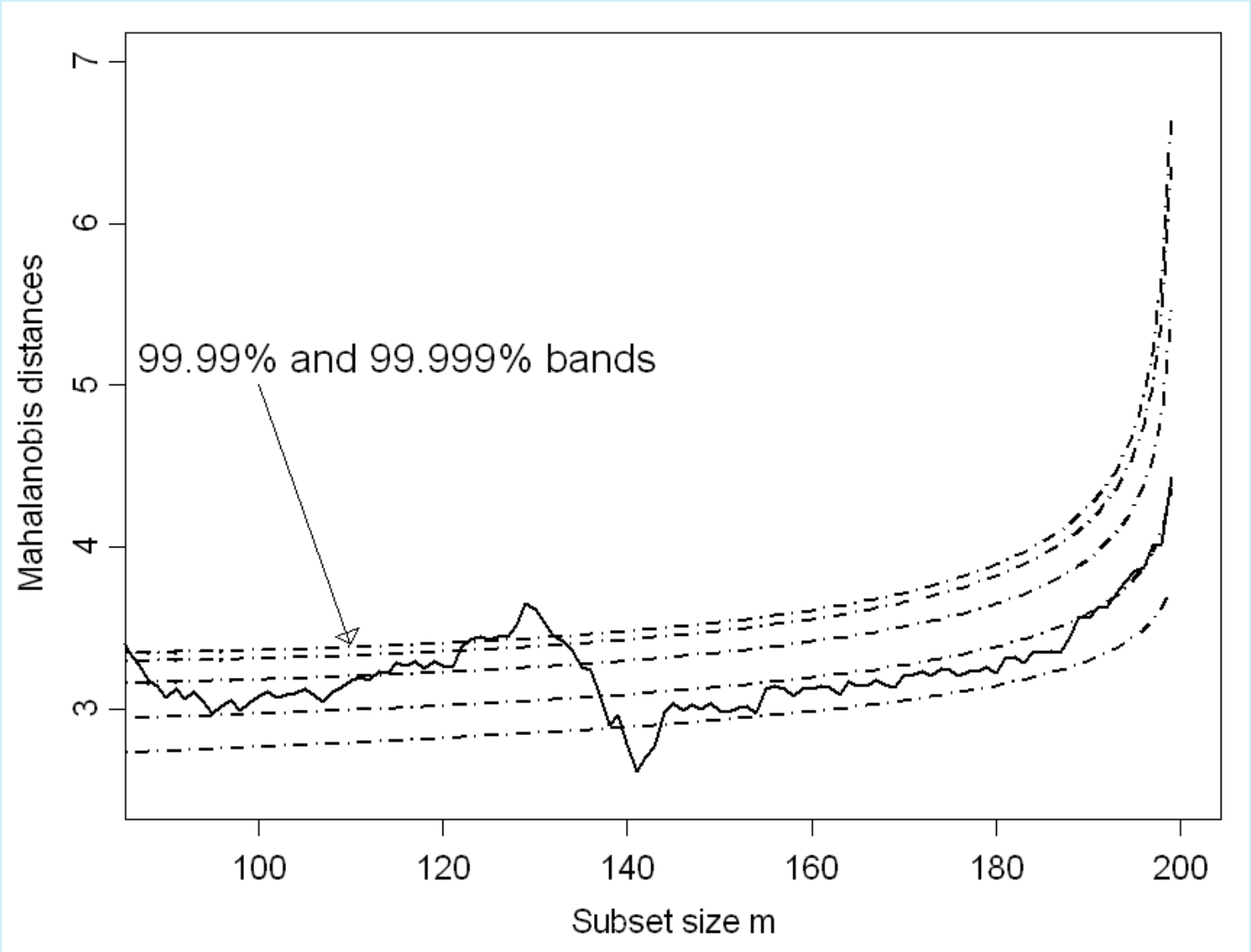
Output from reweighted MCD + Pison correction Real α=0.1095)



Robust MCD distances against traditional MD



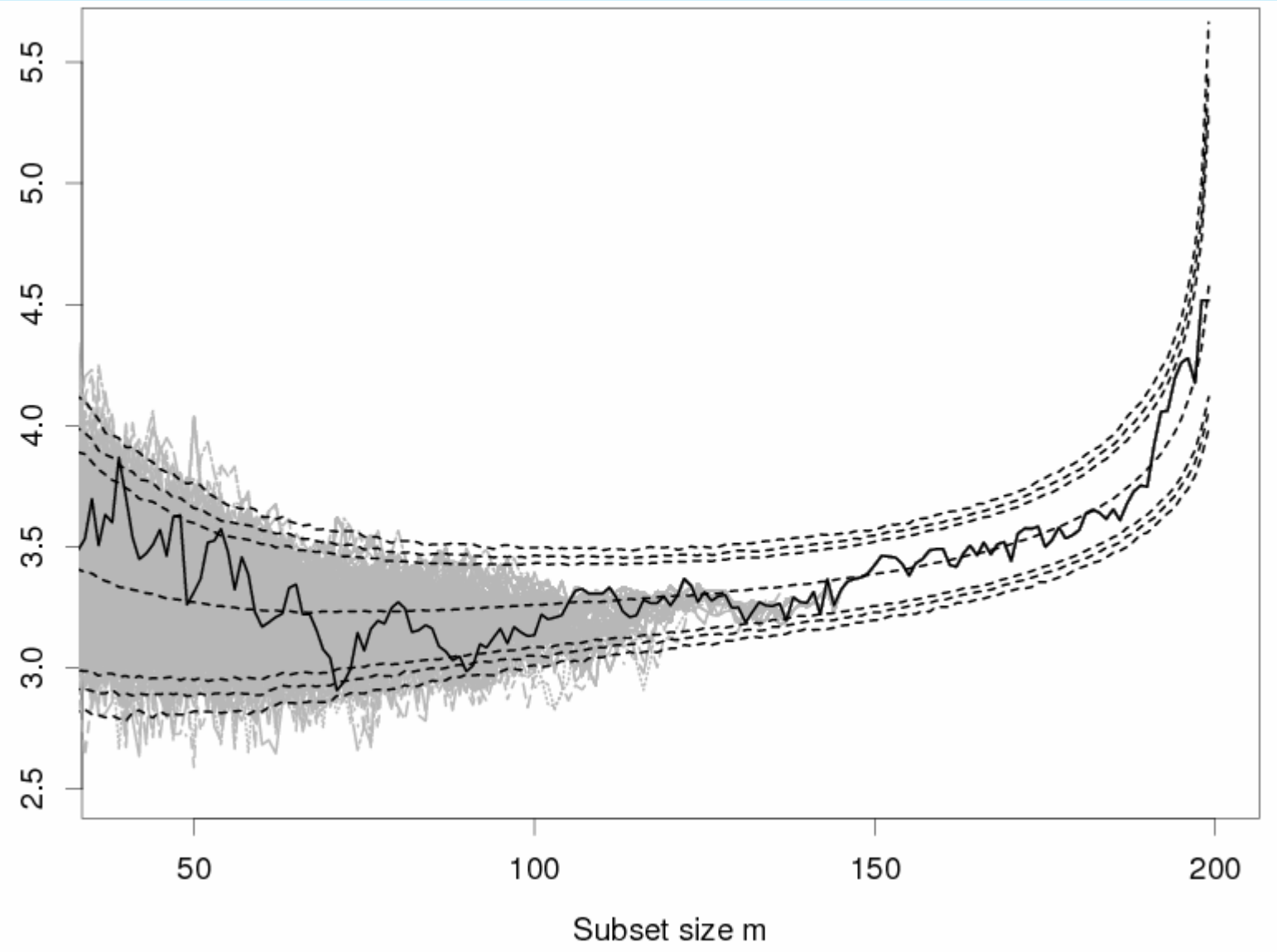
Output from FS



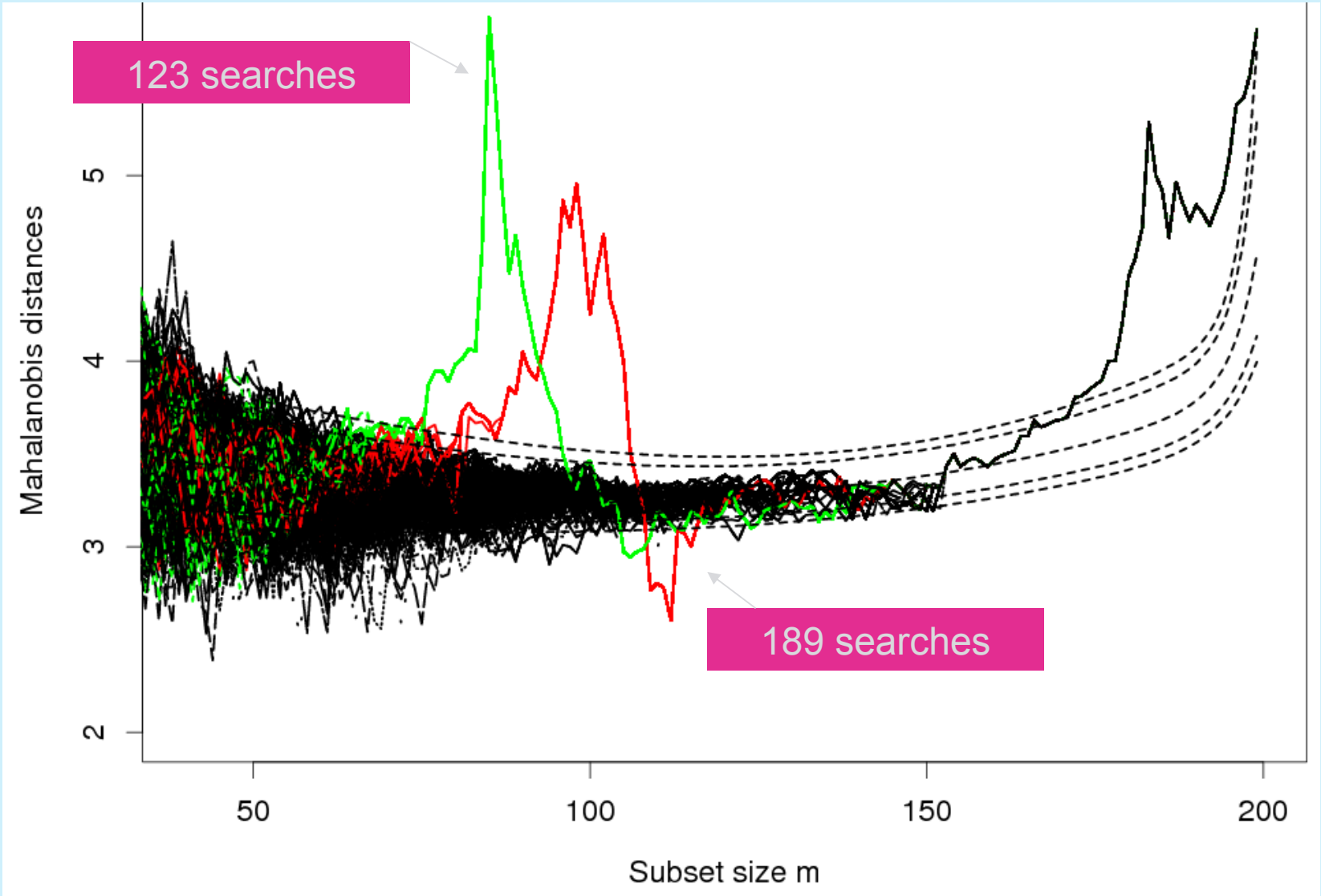
Robust classification through the forward search

The random start approach

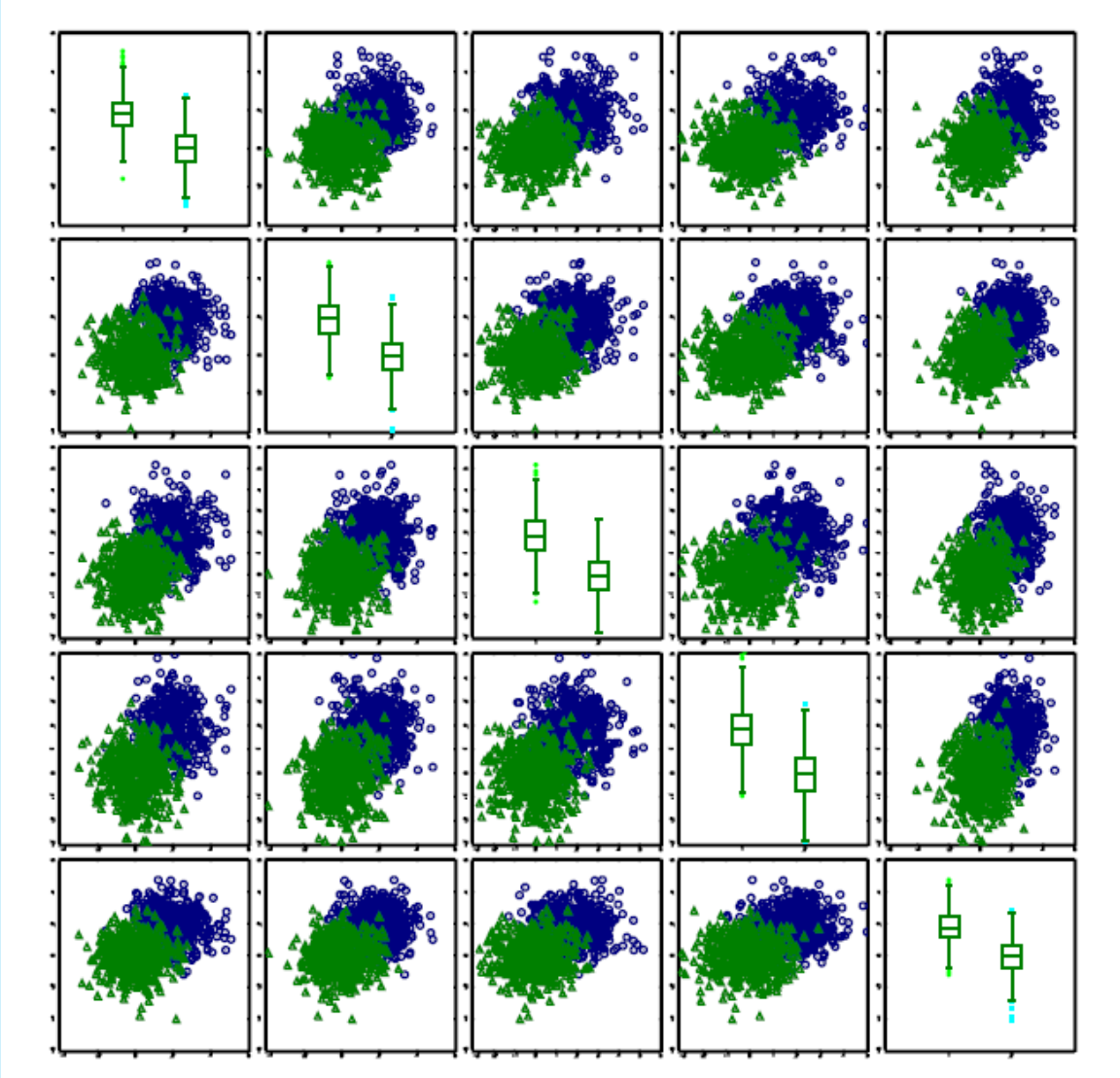
Swiss heads: forward plot of minMD 500 searches with random starting points



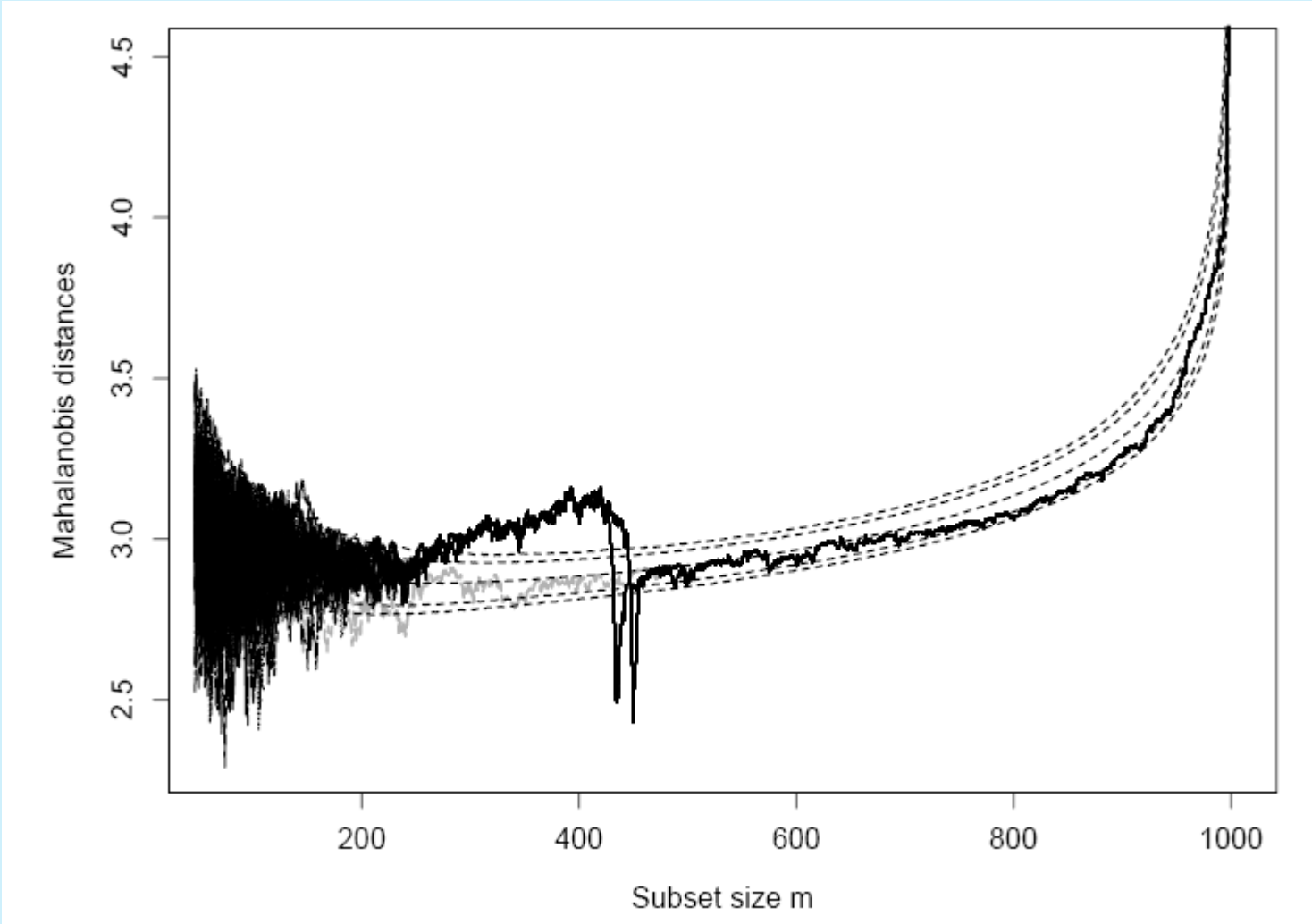
Swiss banknotes: 500 fwd searches (random starts): monitoring of minimum MD



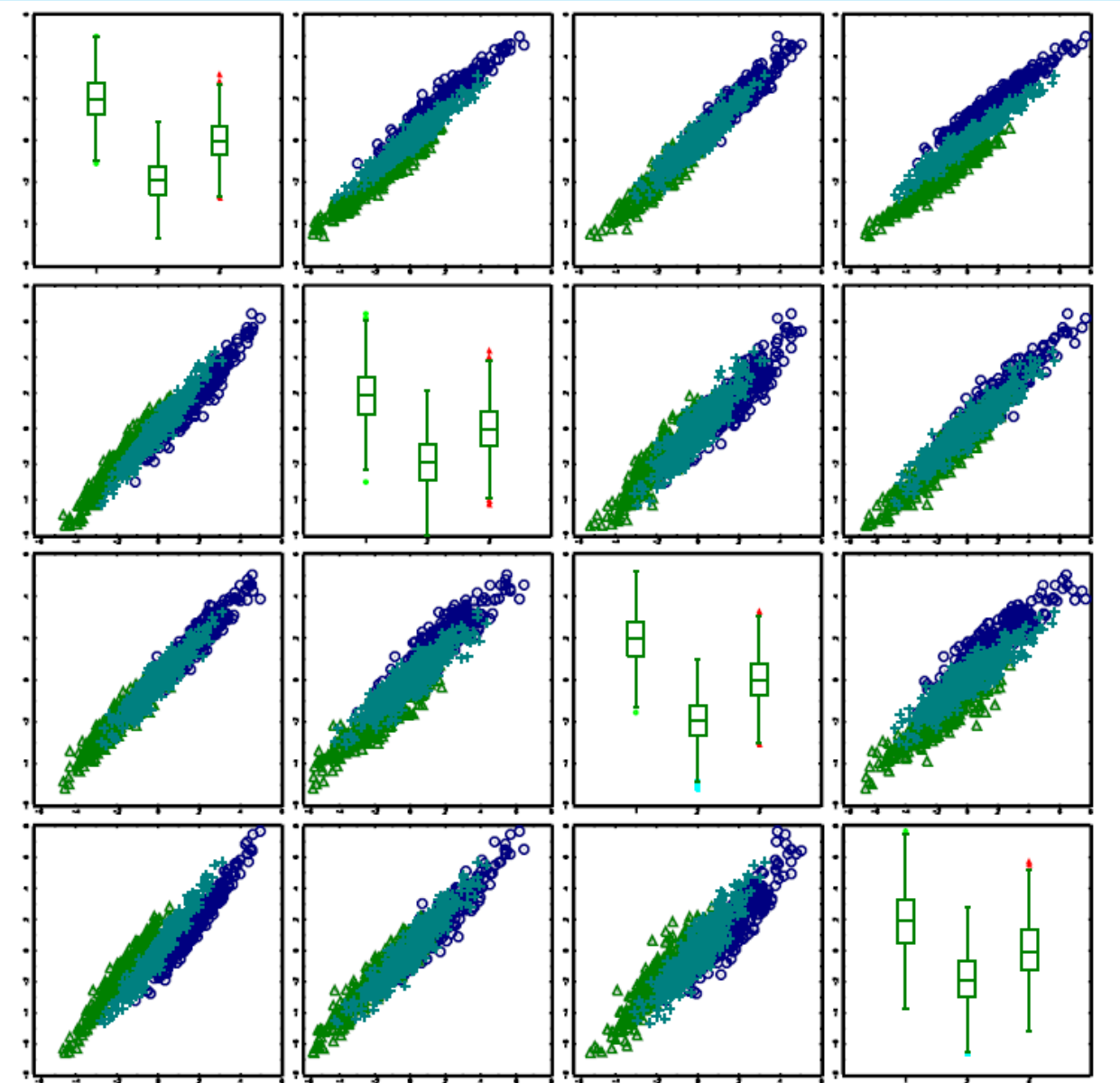
Two clusters of independent normal variables (TC): spm



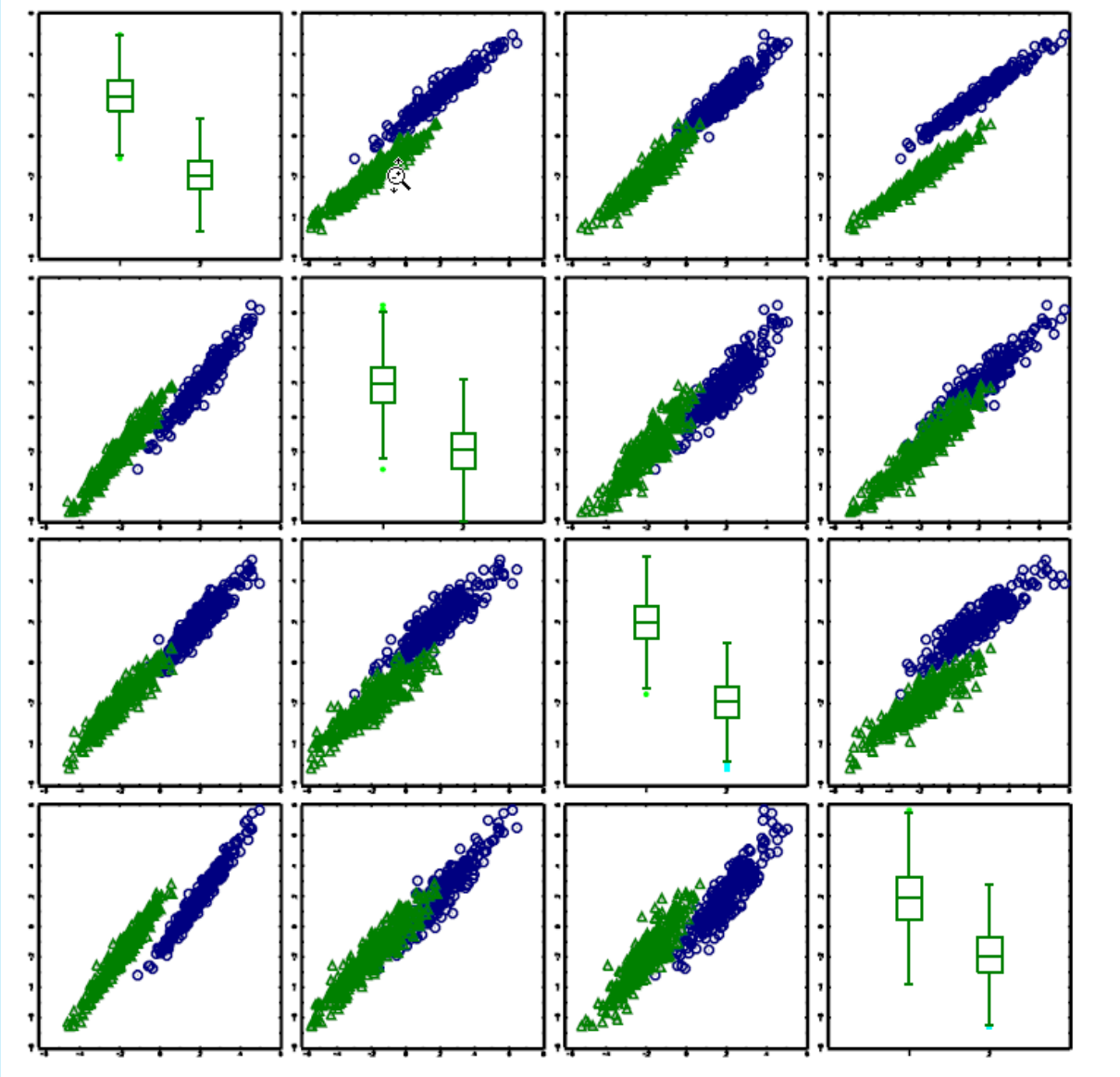
TC: forward plot of Min MD from 200 random starts



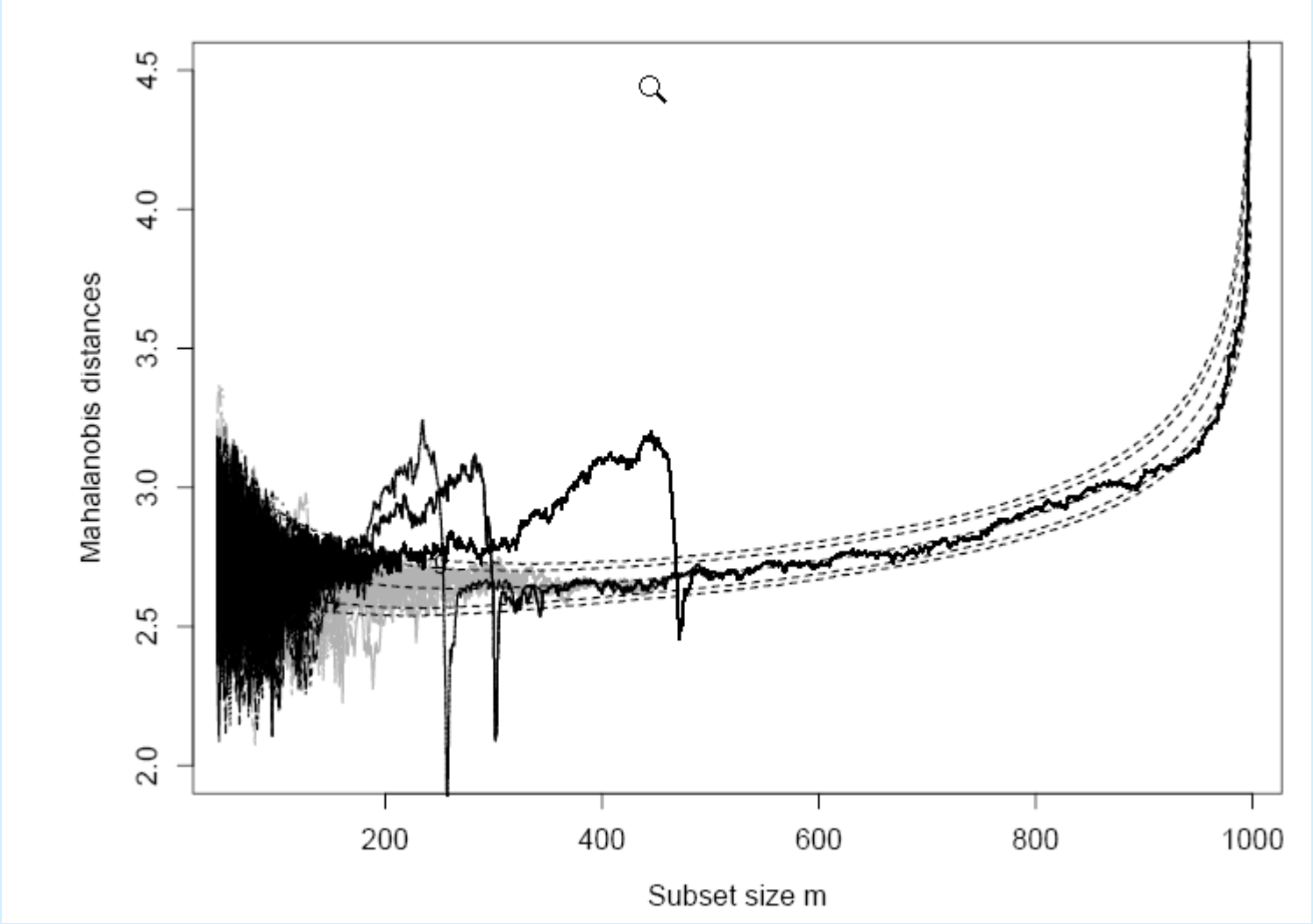
Three clusters of correlated normal variables (3C)



3C: plot of groups 1 and 2



3C: forward plot of min MD from 200 random starts



Classes of robust estimators

Three classes of estimators:

- **Hard (0,1) trimming** (LTS, LMS, MCD, MVE) in which the amount of trimming is determined by the choice of the trimming parameter.
- **Adaptive Hard Trimming**. In the Forward Search (FS), the observations are again hard trimmed, but the amount of trimming is determined by the data, being found adaptively by the search.
- **Soft trimming** (downweighting). M estimation and derived methods (S, MM, tau). rho function ensures that increasingly remote observations have a weight that decreases with distance from the centre.

Decisions which have to be taken when using soft or fixed hard (soft) trimming methods

- The number of subsamples to extract to each of which the model is fitted exactly.
- The maximum number of refining iterations (concentration steps), if any, within each subsample.
- The tolerance for the convergence of the estimate of target function in the refining steps.
- The number of best subsets resulting from the refining steps to be brought to convergence.
- The number of refining iterations for each best subset being brought to convergence.
- The tolerance for the estimate of b in the refining steps for each subset being brought to convergence.
- The tolerance for the estimate of scale in the best subsets.

- Often these choices **are not well documented** in software

The philosophy of monitoring

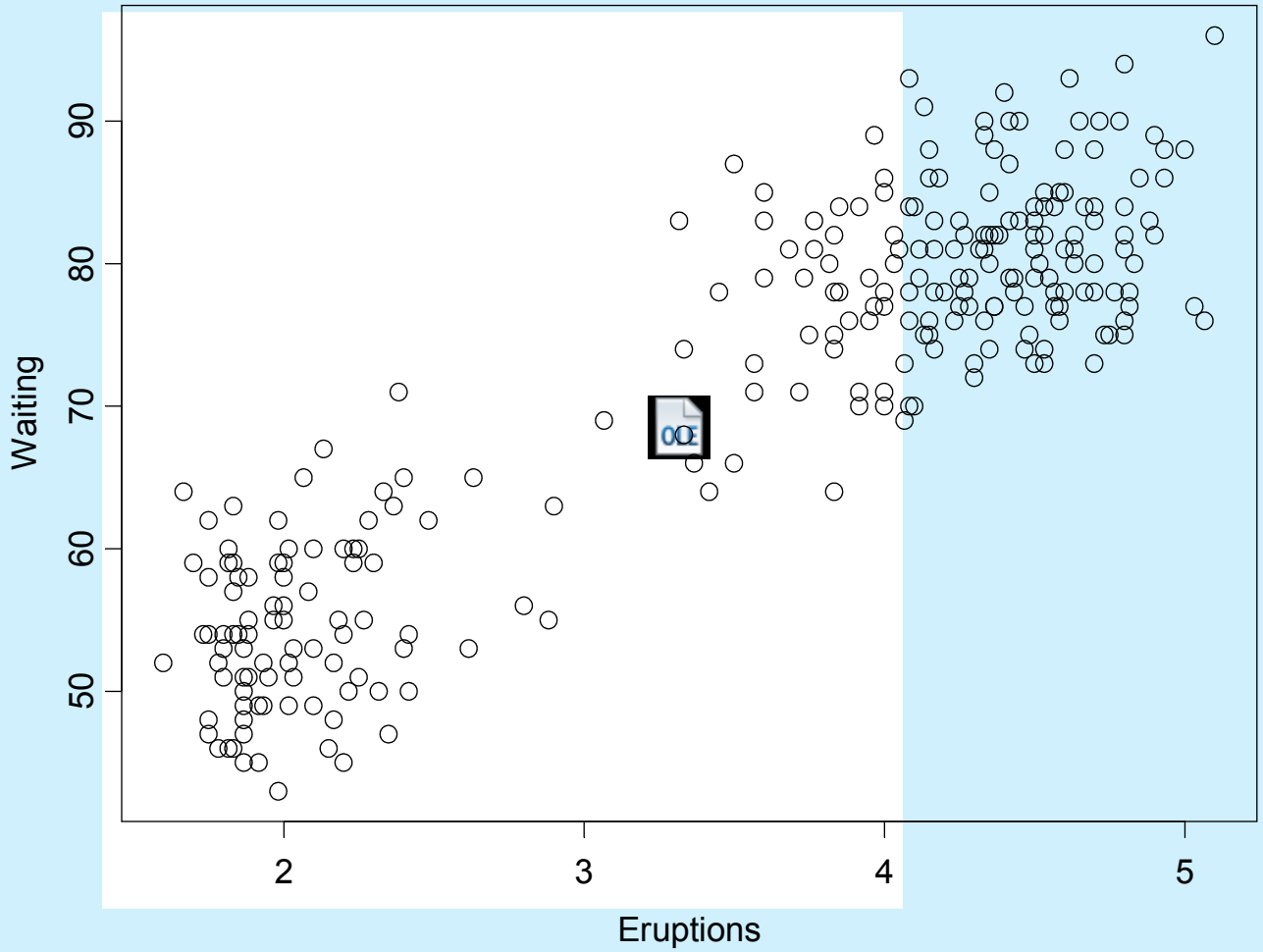
- One reason for the excellent performance of the Forward Search is the **adaptive choice** of the trimming parameter
- Extension: **MONITOR** the behaviour of robust procedures over a range of values of this parameter
- Monitoring also helps with the **choice among robust methods and the decisions that have to be made before data analysis**
- **These decisions are another major disincentive to the routine use of standard robust methods**

Geyser data

2 variables

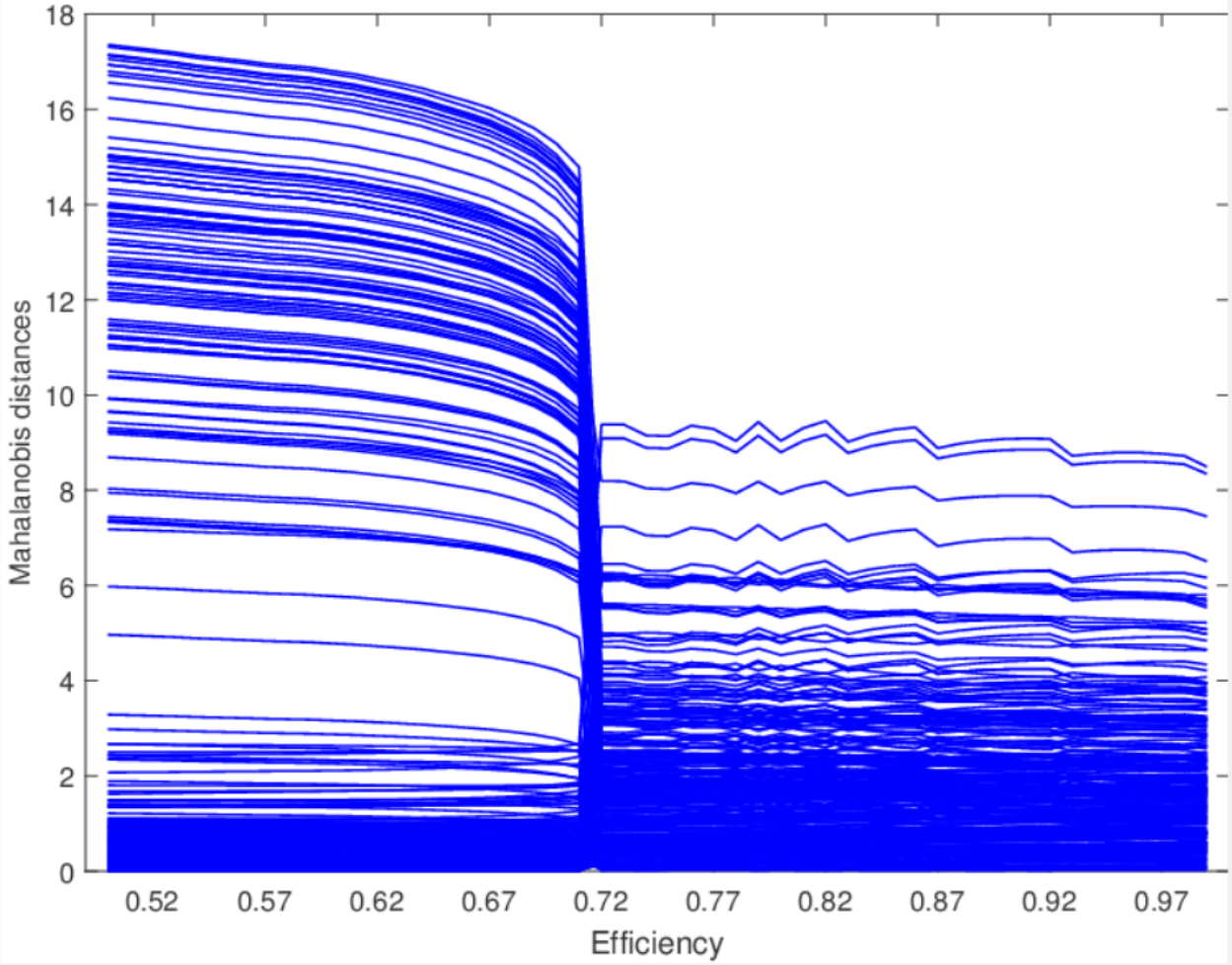
n=272

Geyser data (n=272)



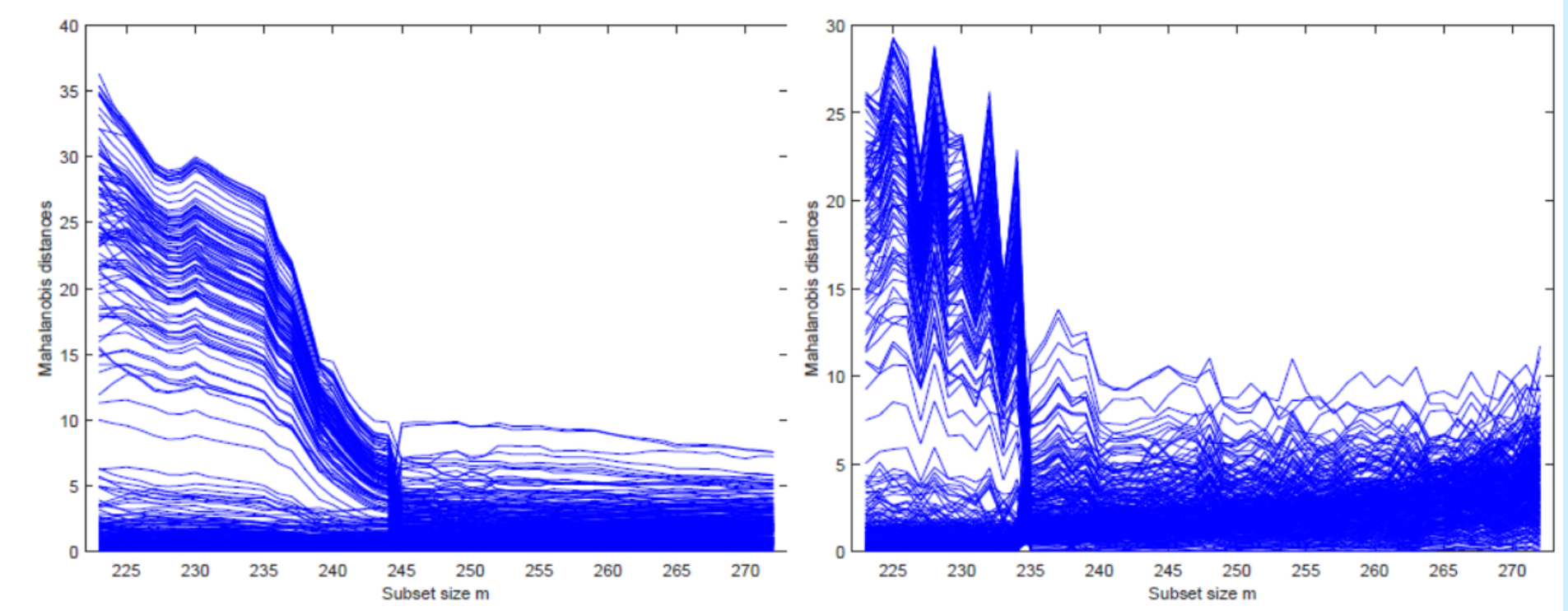
- Big cluster ≈ 175 obs
- Small cluster ≈ 97 obs

Robust Mahalanobis distances for MM estimation as a function of eff from 0.5 to 0.99



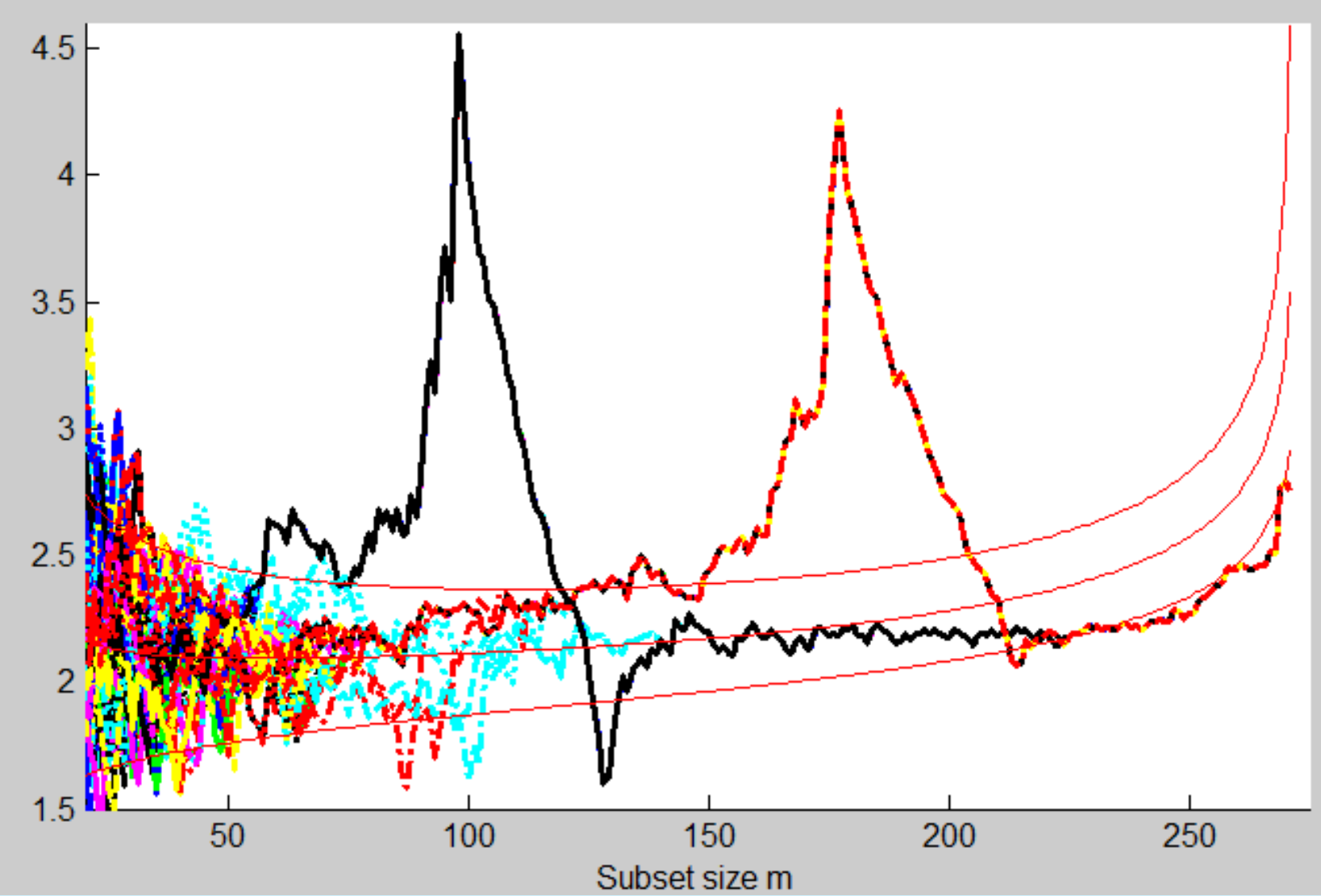
For efficiencies less than 0.71, the plot reveals the observations from the smaller cluster as outliers

Robust Mahalanobis distances for MCD (left) and MVE (right) as a function of subset size

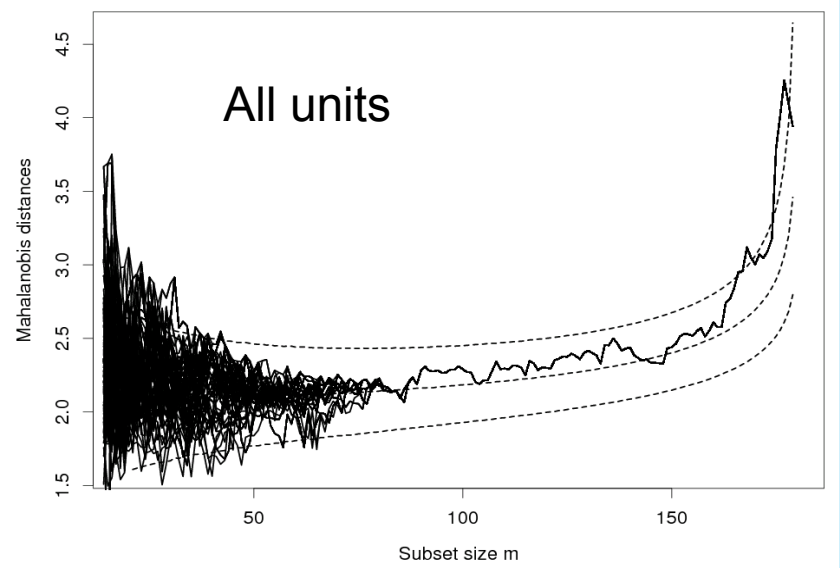


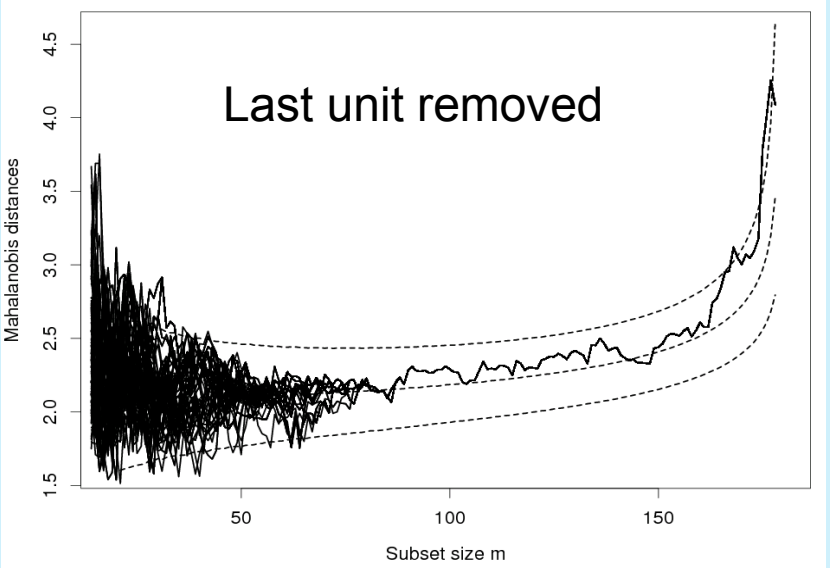
- MCD: definite indication of a break in the structure, here at a subset size of 244

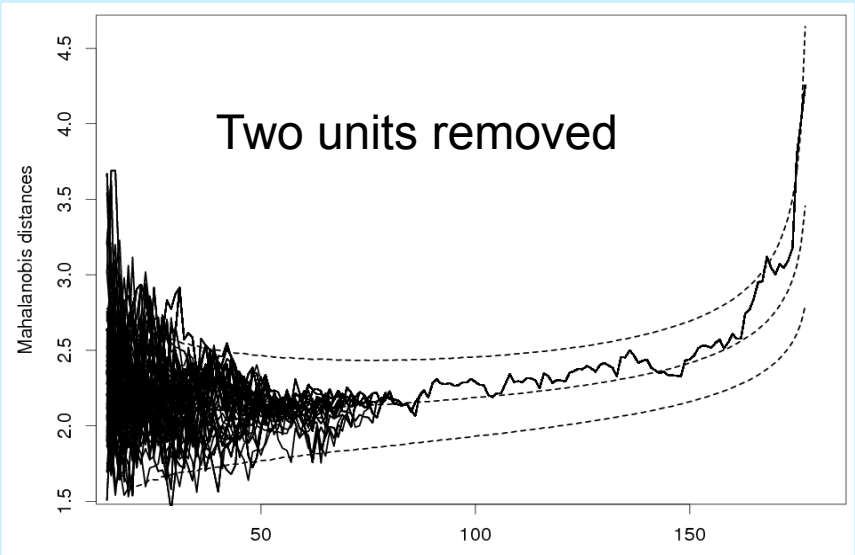
A random start forward search analysis of the Old Faithful data: monitoring of min. Mahalanobis distance outside subset

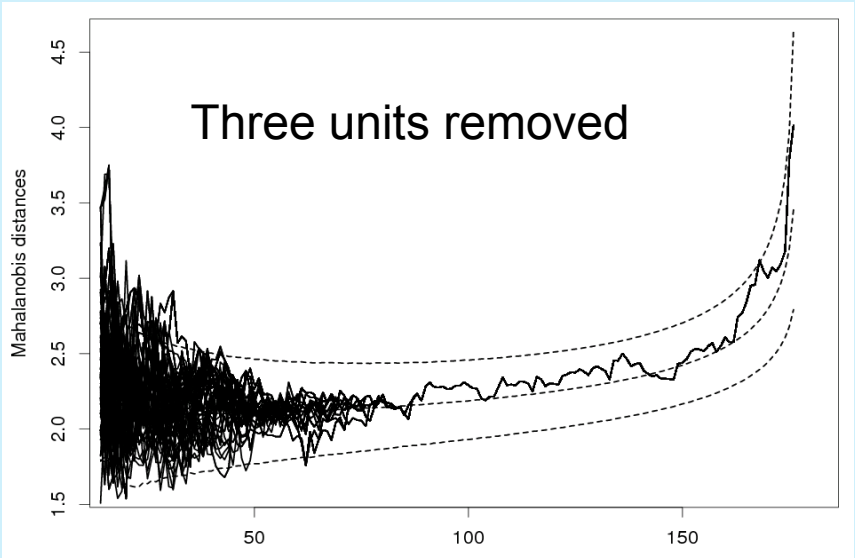


Further analysis within each cluster to find the number of units belonging to each group

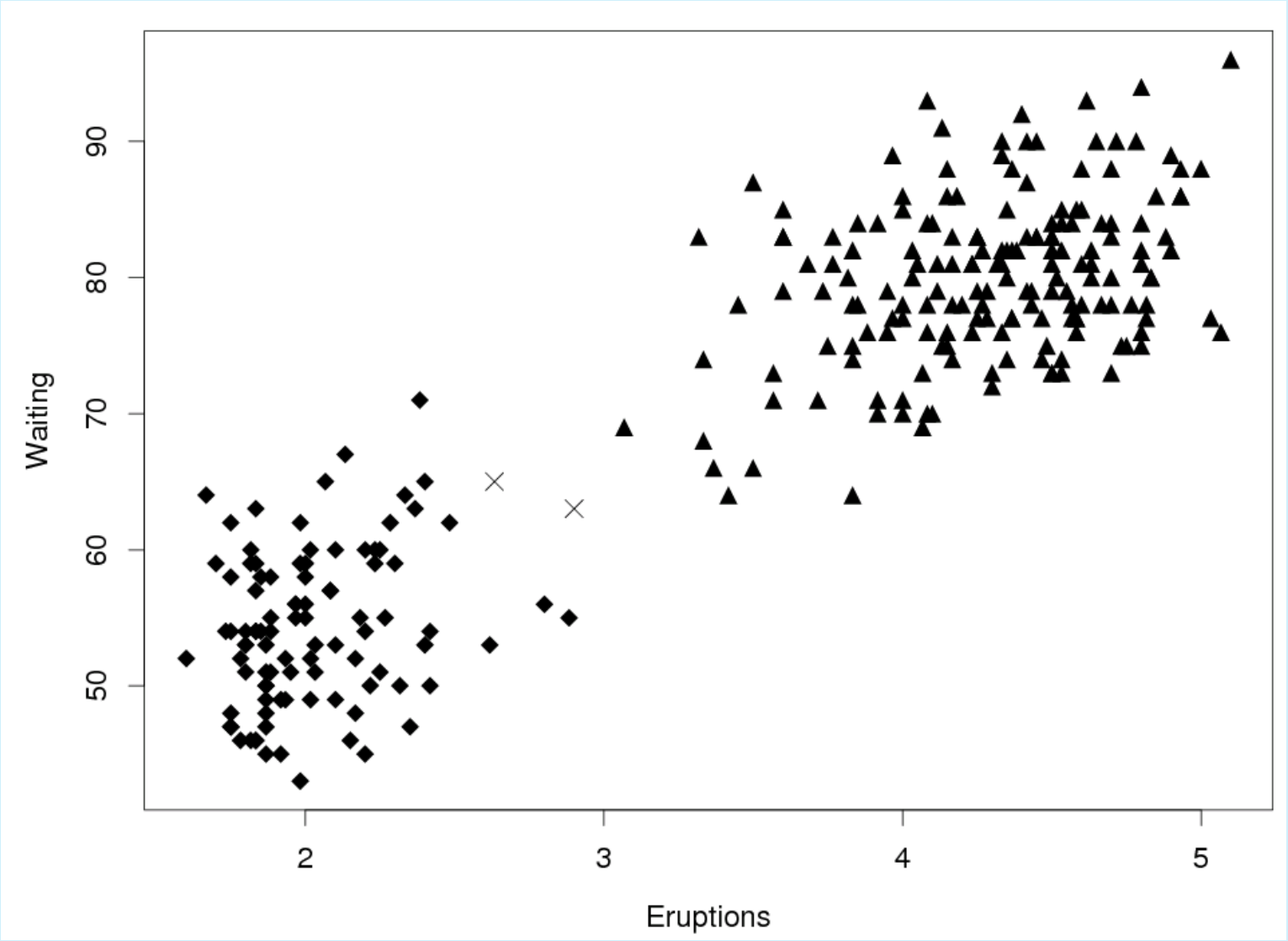




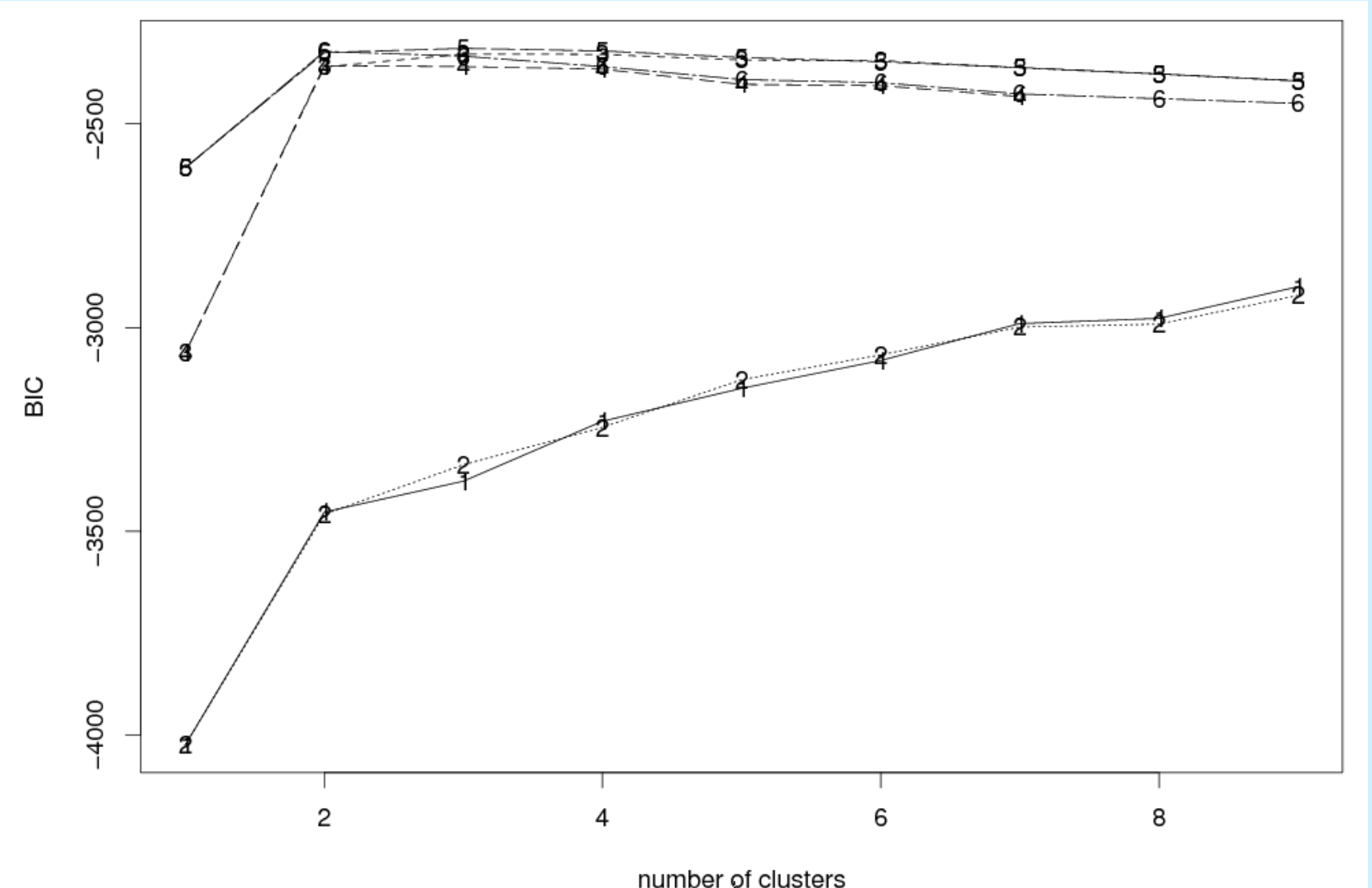




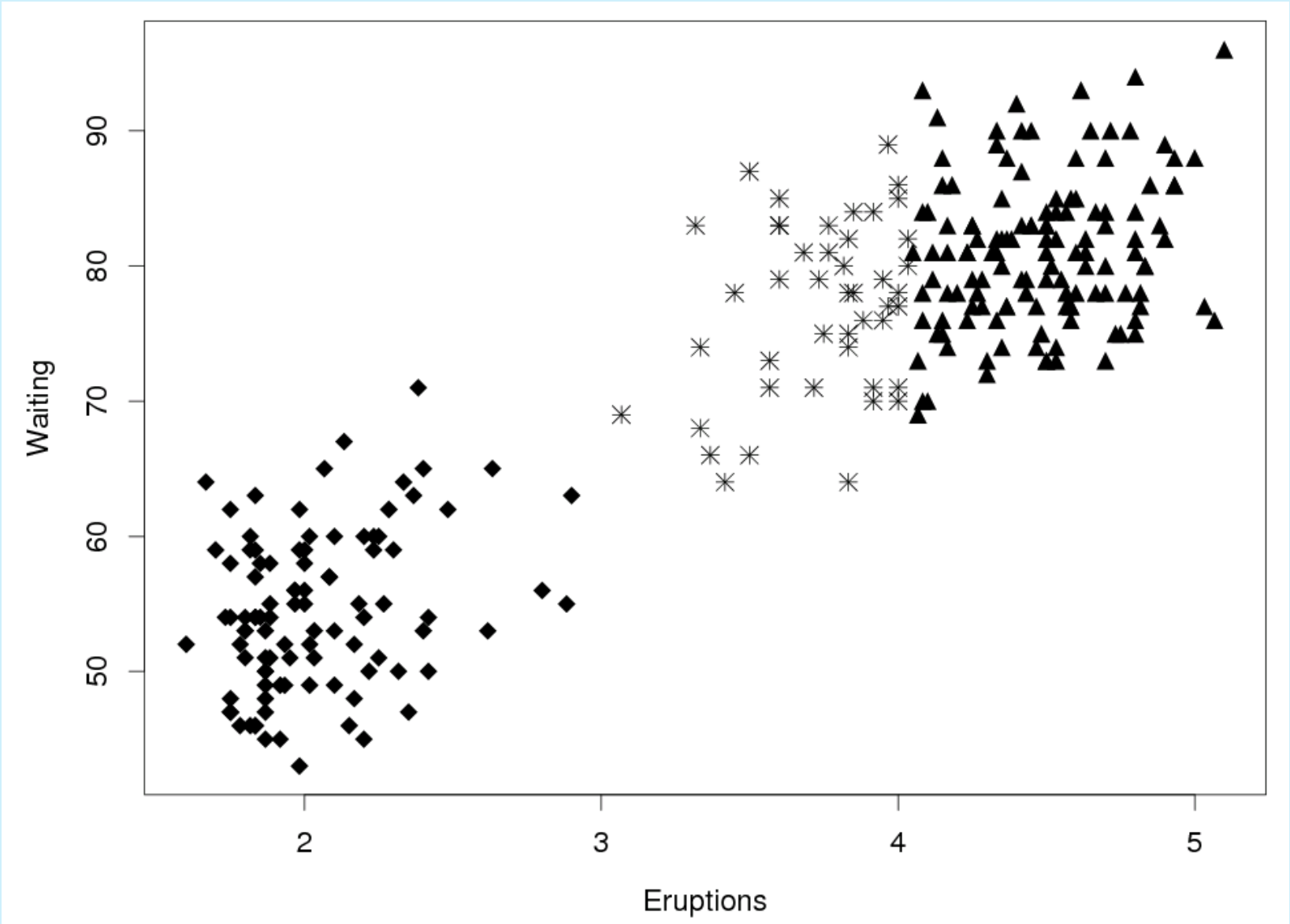
Final classification from the FS



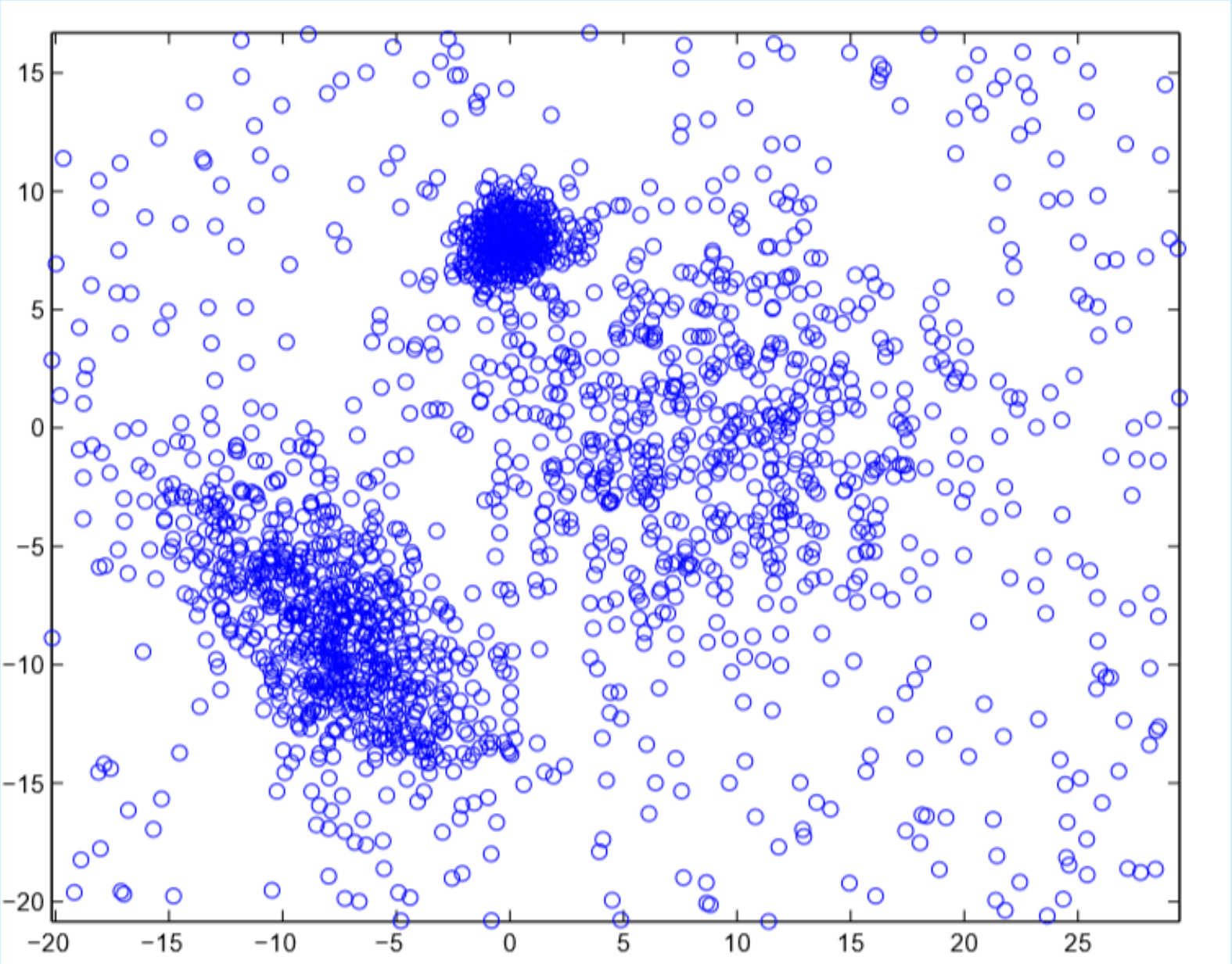
Comparison with Mclust



Final classification from MCLUST

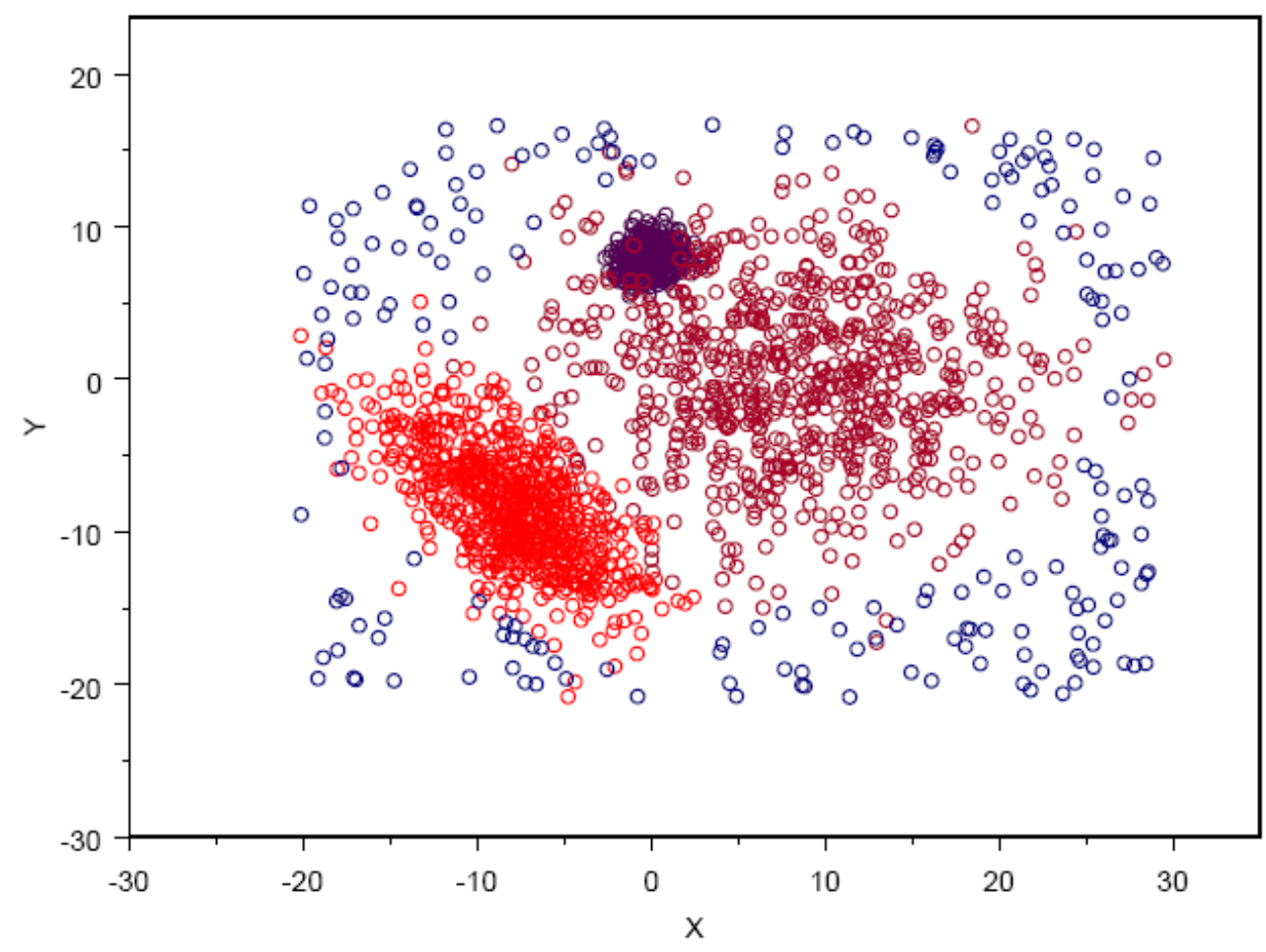


Another example: M5 dataset



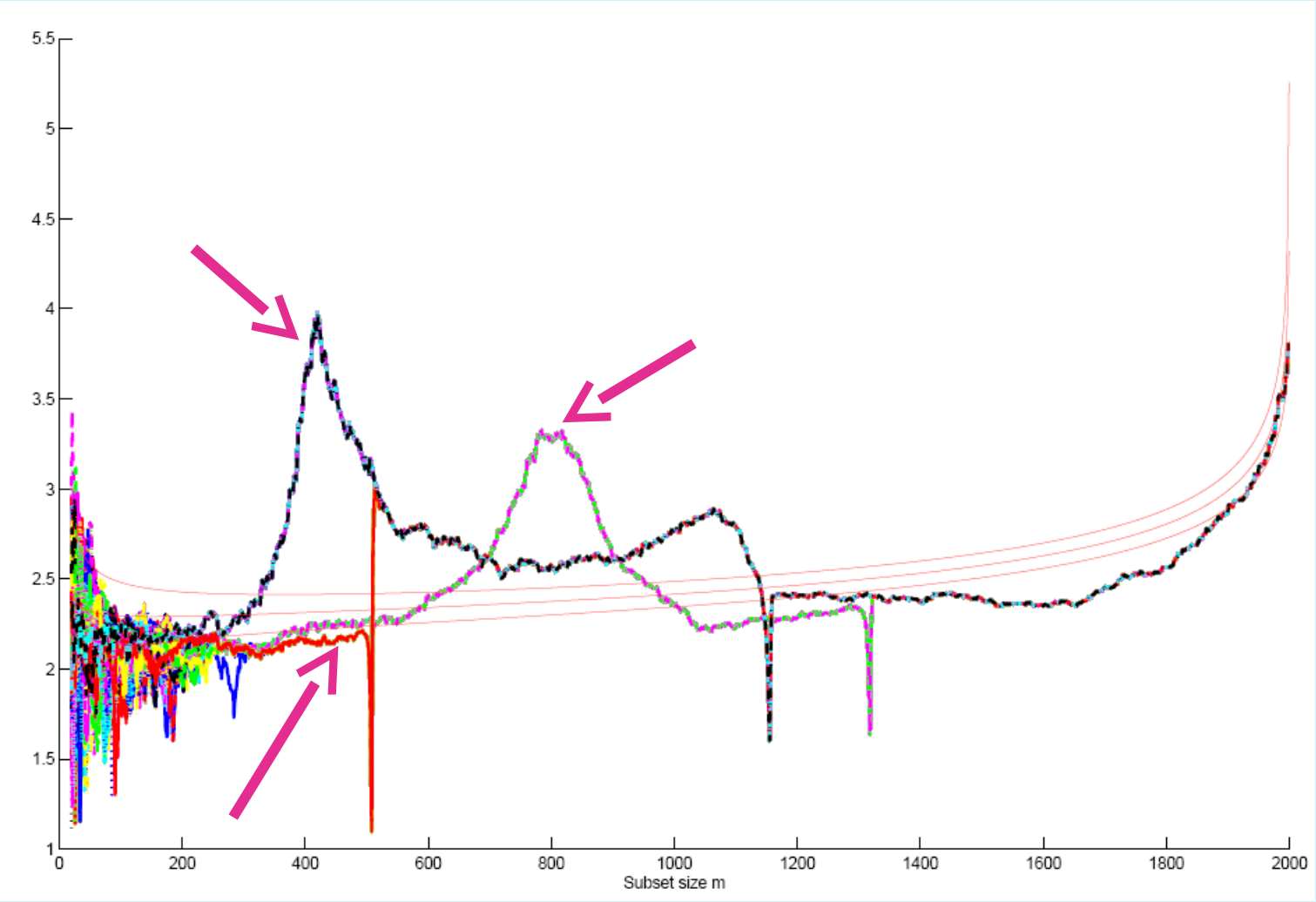
Another example: M5 dataset

- $n = 2000$ observations: 1800 "good" data points, simulated on v = 2 (normal) variables, + 200 outliers
- **Three groups** with different scales



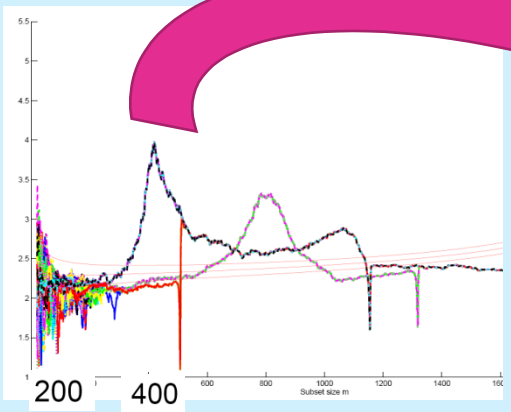
Radial outliers around the groups; two groups strongly overlap

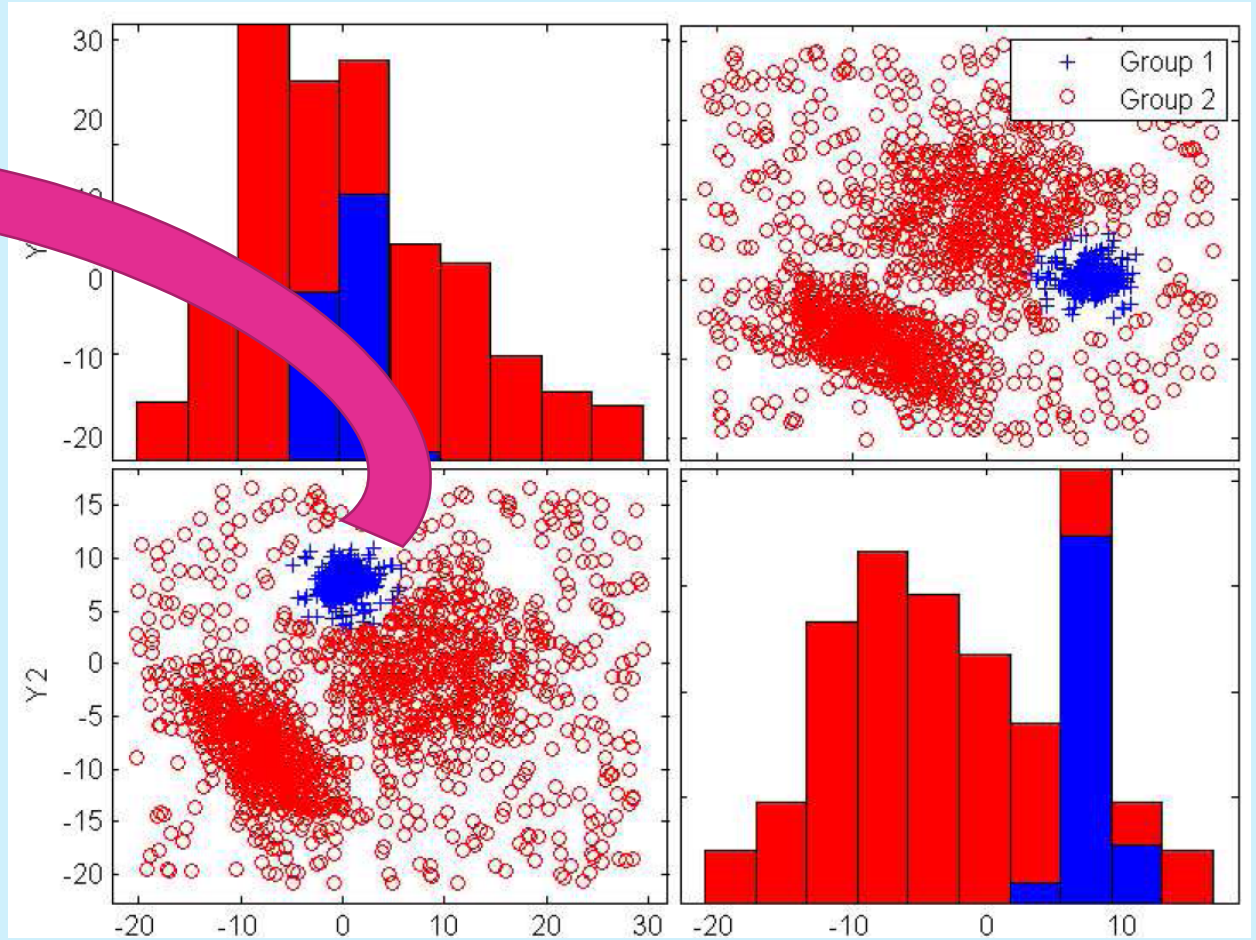
M5: 500 fwd searches (random starts): monitoring of minMD



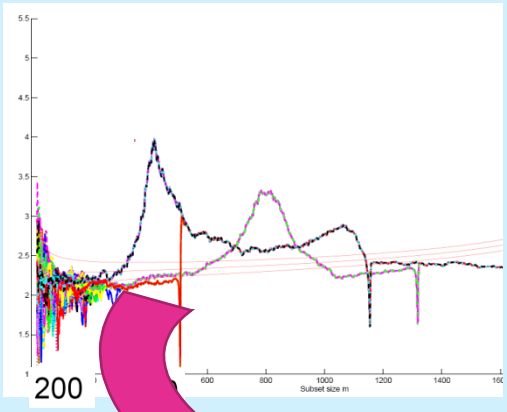
3 different trajectories

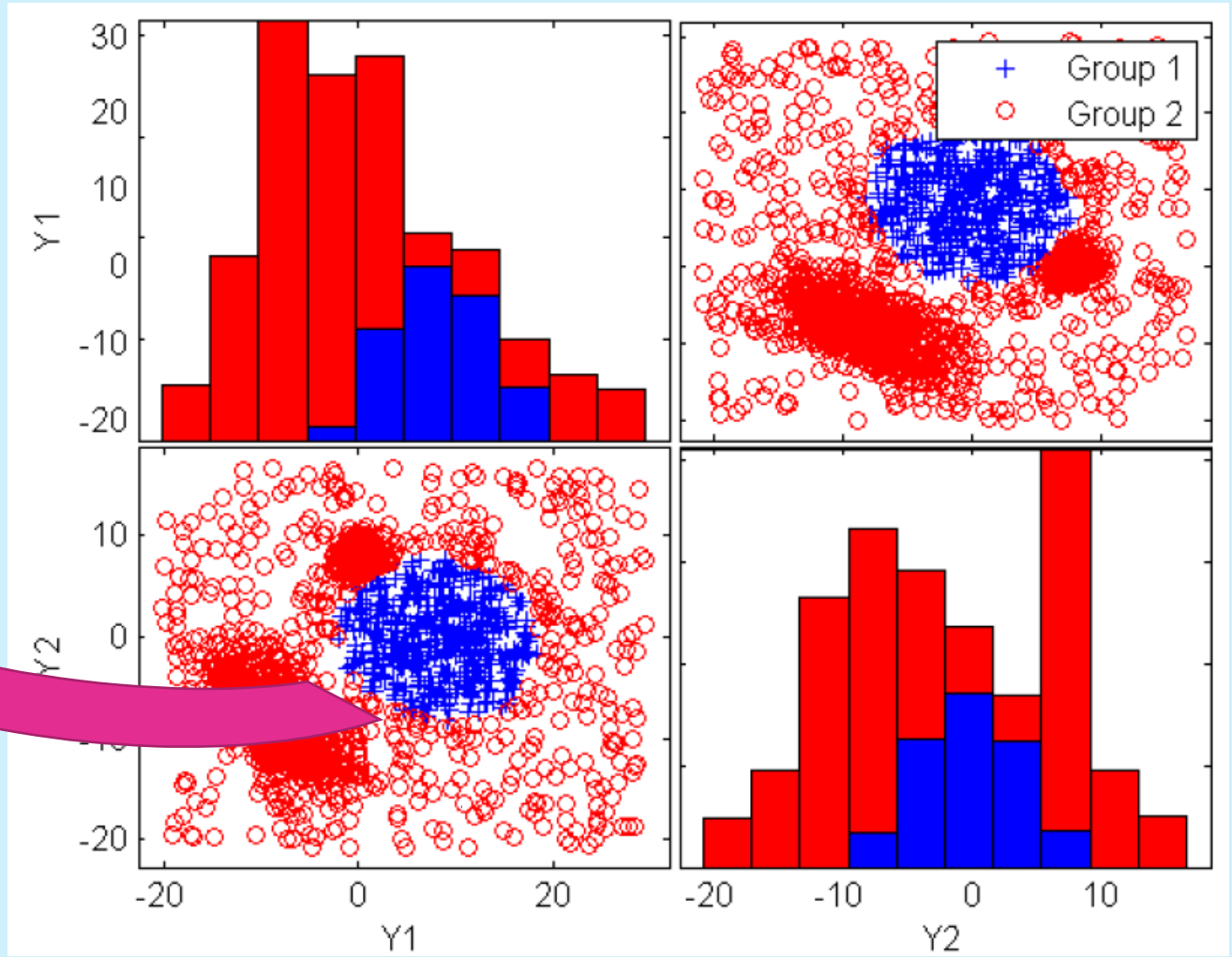
We can interrogate the FS at selected steps: step m=420



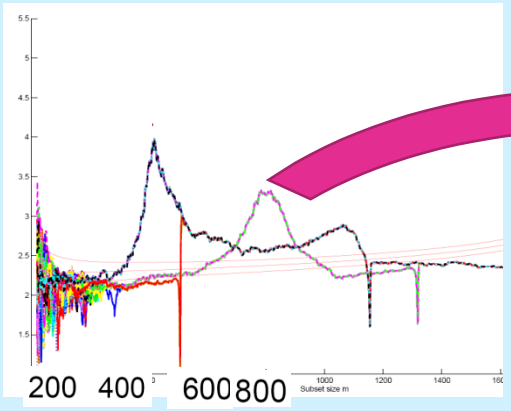


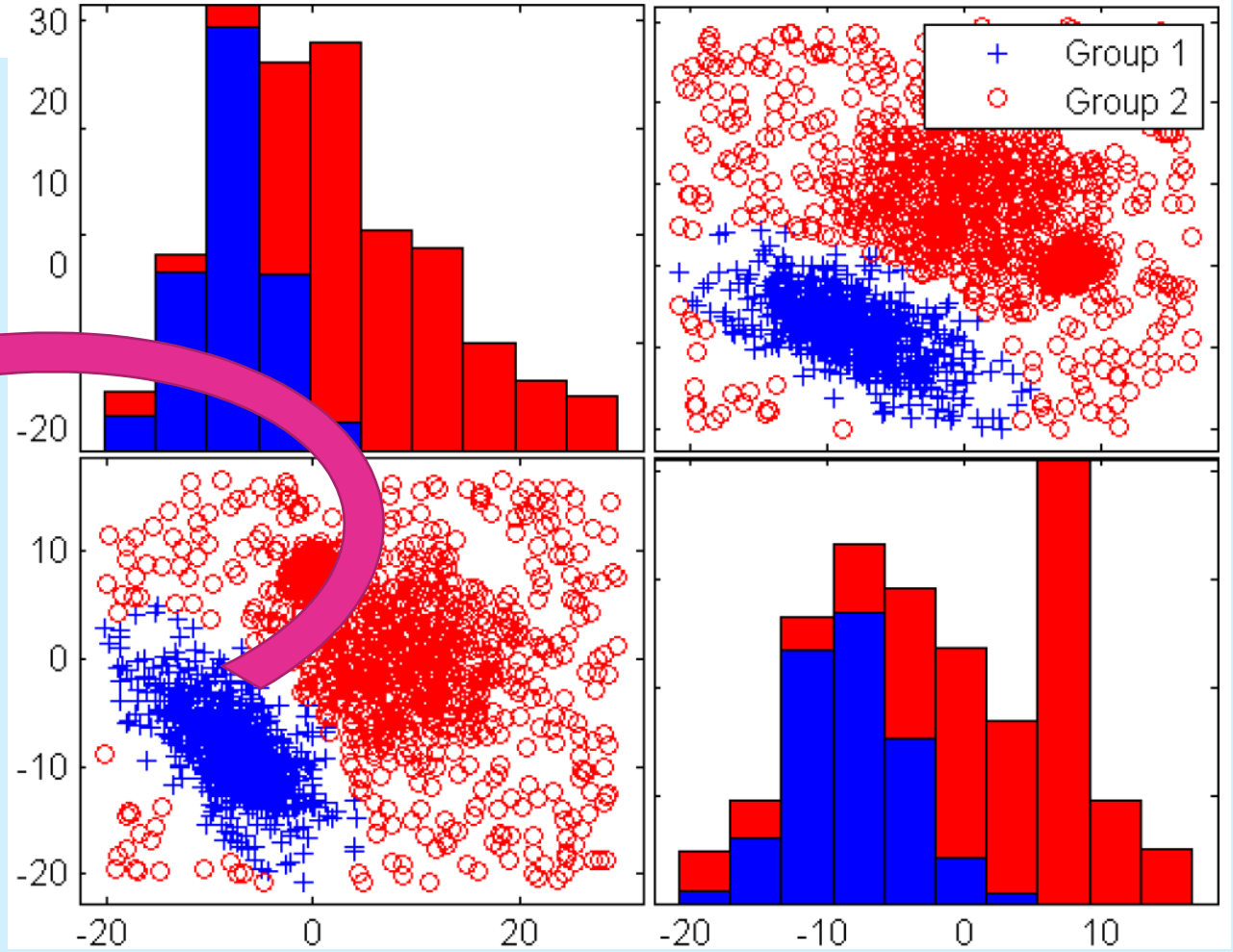
Step m=490



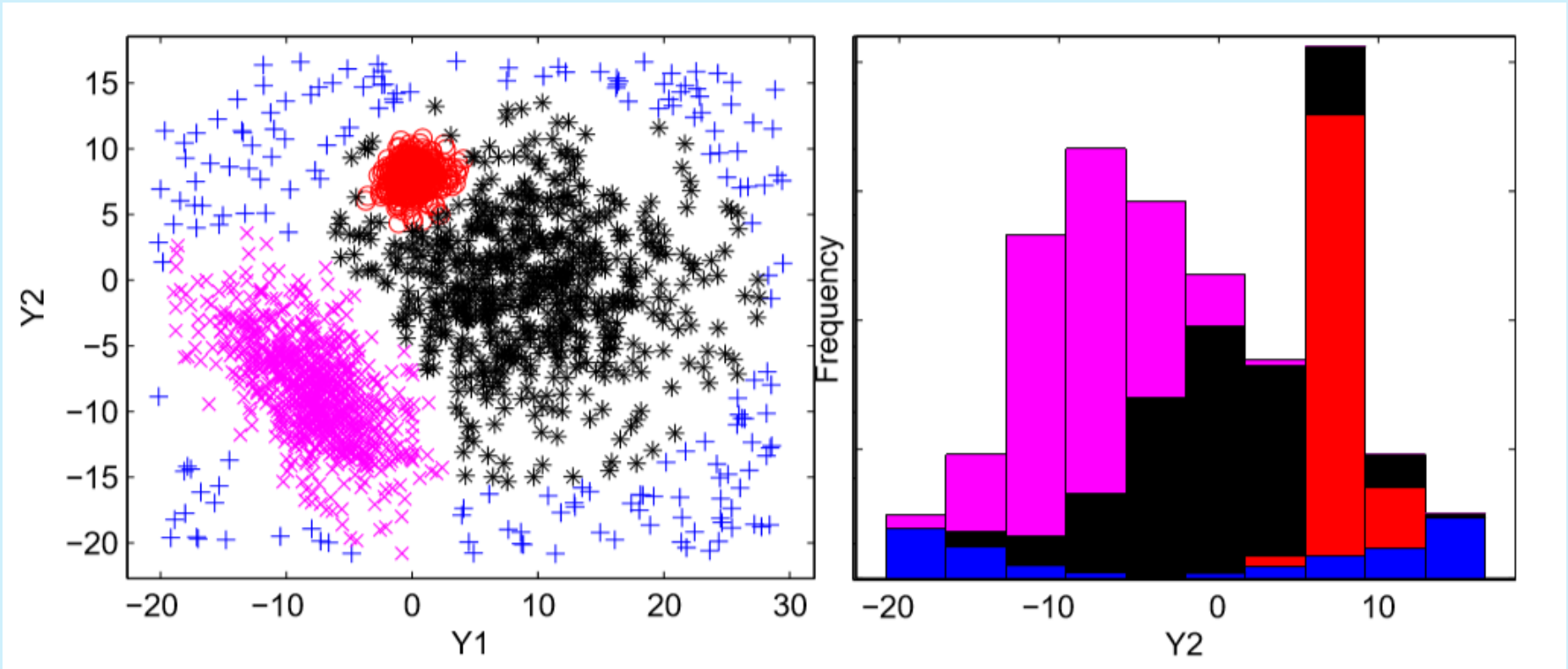


Step m=780


200
400
600
800
Subset size m


Group 1
Group 2
30
20
10
0
-20
10
0
-10
-20
-20
0
20
-20
-10
0
10

Final classification



Comparison with leading robust clustering routines

- **TCLUST**: mainly developed in Valladolid (Garcia-Escudero et al., *Ann. Stat.*, 2008; *ADAC*, 2010; Statistics and Computing 2012)

TCLUST

Model-based robust clustering approach:

$$f(y) = \left[\prod_{j=1}^{k} \prod_{i \in R_j} \pi_j f(y_i, \mu_j, \Sigma_j) \right] \left[\prod_{i \notin R} g_i(y_i) \right] = [L_R(y)]\,[L_O(y)]$$

- $f(y)$ is the v-variate **normal density** for $y = \{y_1, \ldots, y_n\}$
- $R_j, \ldots, R_k$ are the **"regular" groups**, $R \bigcup_{j=1}^{k} R_j$, and π_j is the j-th group weight
- $g_i(y_i)$ is the density of the i-th contaminated observation
- $\#R = n - \lfloor n\alpha \rfloor$ is the number of **"good" observations**
- $\lfloor n\alpha \rfloor = n - \#R$ is the number of **contaminated observations**

Issues in Tclust

- **Heteroscedastic model**: we need **constraints** on the covariances of the different groups

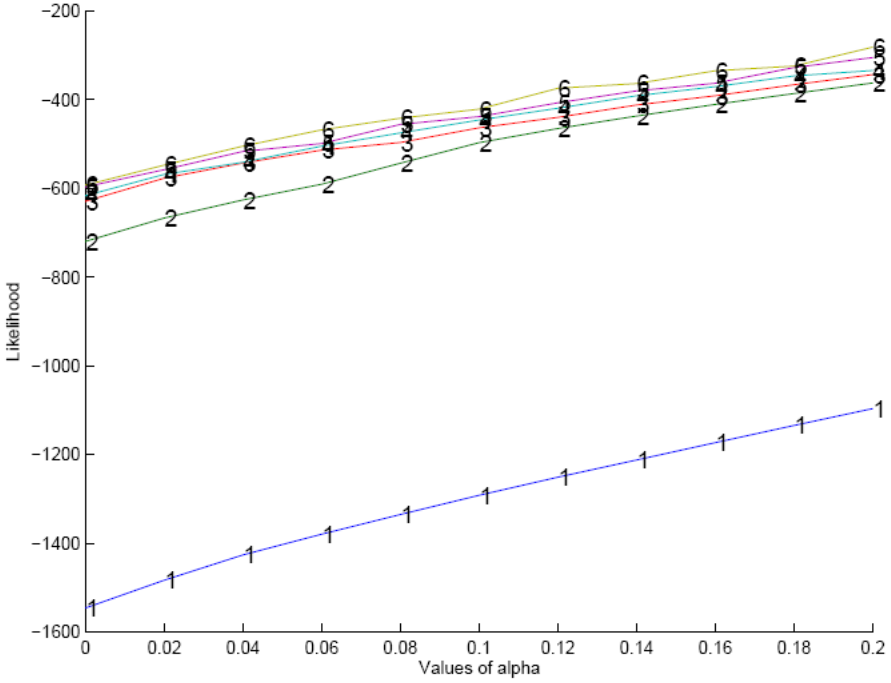
$$\frac{\max_{j=1,\ldots,k} \max_{l=1,\ldots,v} \lambda_l(\Sigma_j)}{\min_{j=1,\ldots,k} \min_{l=1,\ldots,v} \lambda_l(\Sigma_j)} \leq c$$

- Three crucial aspects:
- α = trimming proportion
- k = number of groups
- c = restriction factor

State of the art for the choice of k and α

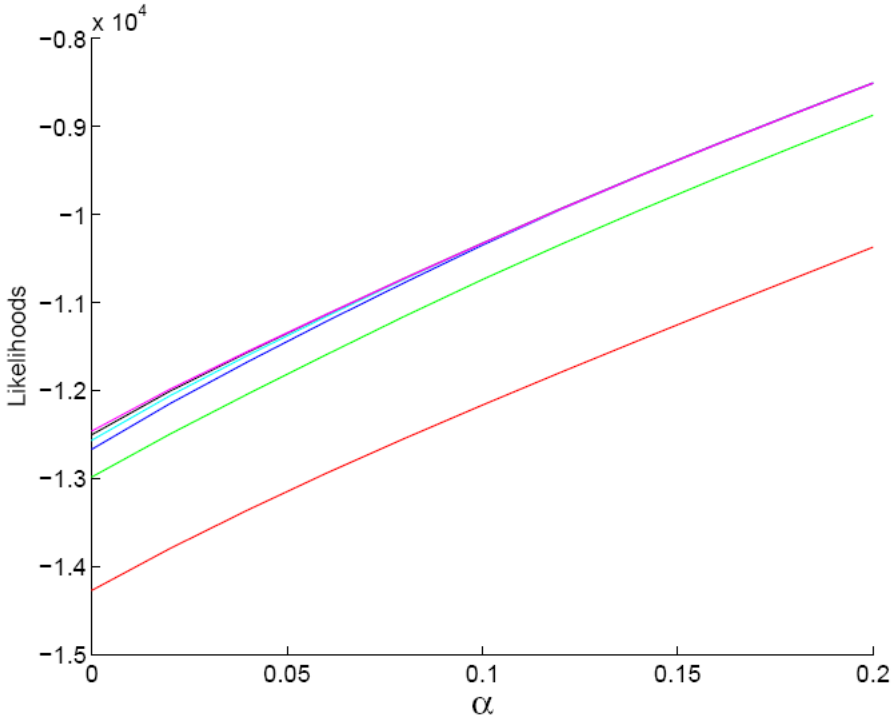
- Classification trimmed likelihood (CTL) curves

- Swiss banknotes



Suggests k=2 and α=0.1

- M5 datasets



k=3 but α not clear

All routines for robust estimators are implemented in the MATLAB toolbox FSDA downloadable from http://www.riani.it/MATLAB or from http://fsda.jrc.ec.europa.eu