# Regression Diagnostics and the Forward Search 5. Several Multivariate Populations

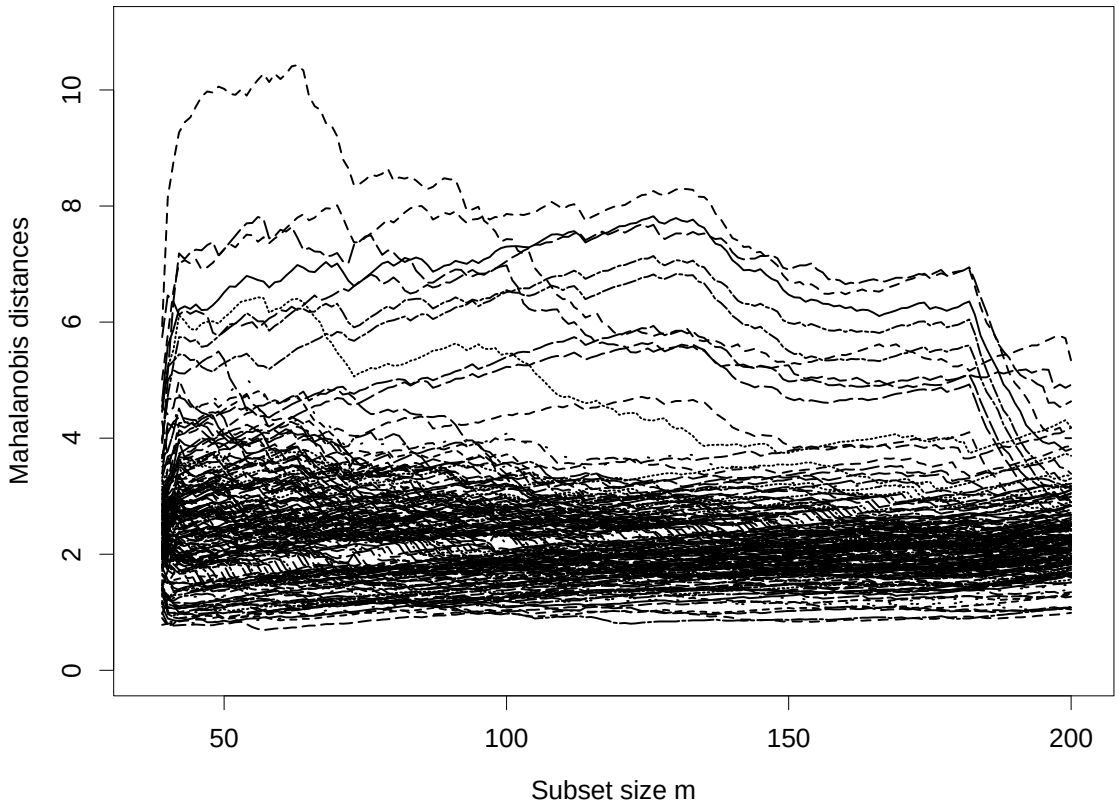
Anthony Atkinson, LSE

# The Forward Search: Several Populations

- Estimation for a single population often involves minimizing a single convex function
- One Forward Search likewise reveals the structure of a sample from a single population (+ outliers)
- For several populations, several functions have to be minimized
- Several searches are often needed to elucidate the structure when there are several populations

# Swiss Banknote Data Again

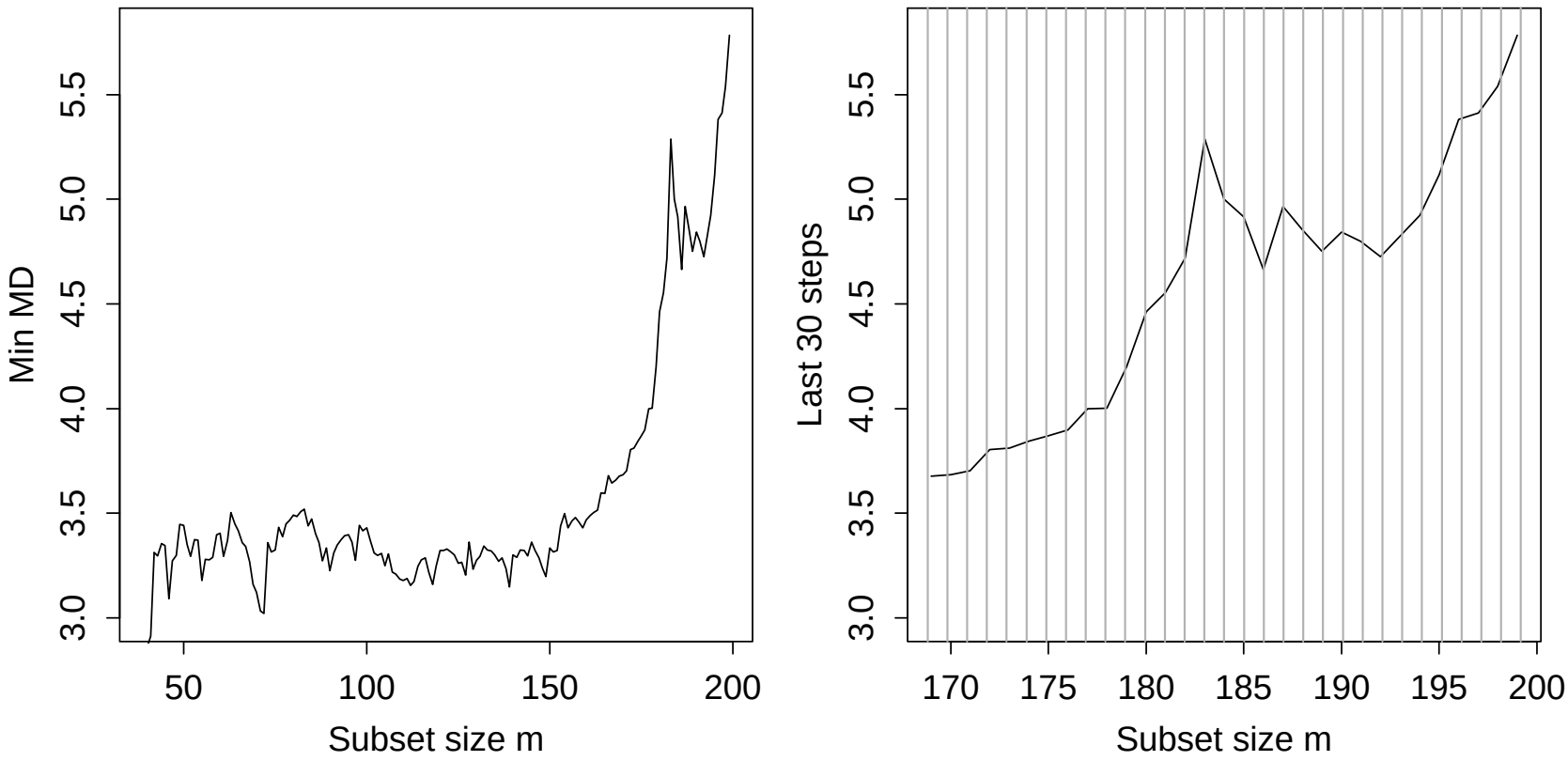
- There are (to recall) 200 observations on six variables
- All notes have been withdrawn from circulation and classified by an expert
- 100 notes are “genuine”, 100 “forgeries”
- But the notes may be misclassified
- There may be more than one forger
- What we see in the Forward Search depends where we start




Swiss bank notes: forward plot of scaled Mahalanobis distances from the search starting with a subset of units from both groups. There seem to be many outliers, but the group structure is not clear

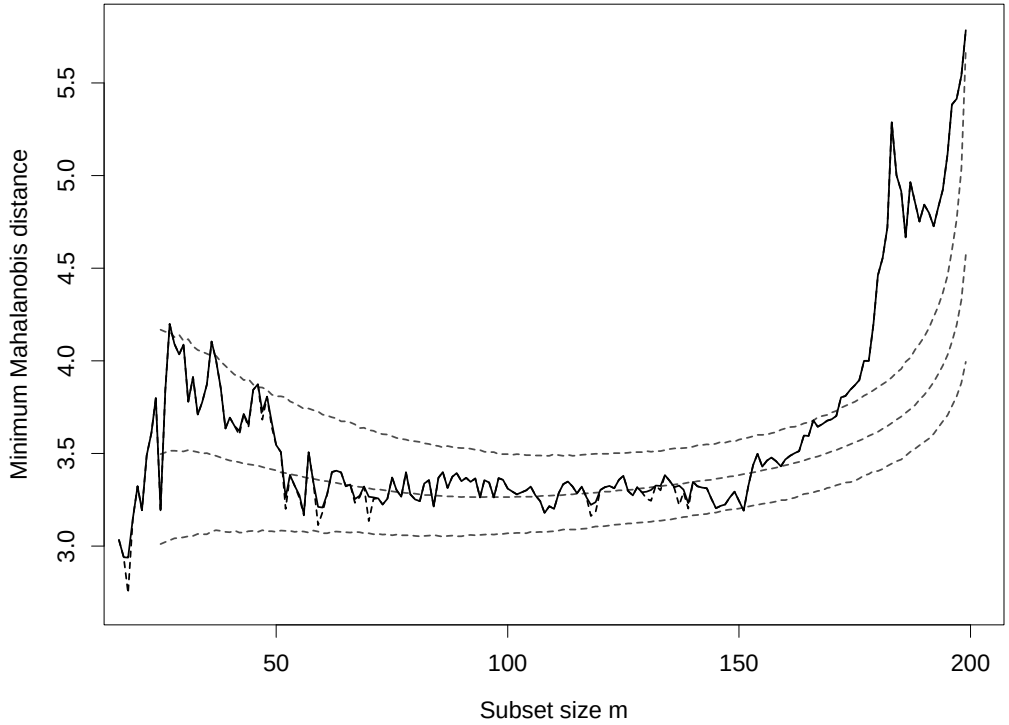
# Swiss Banknote Data 2

To determine whether there really are any outliers, we can look at the forward plot of minimum distances of units not in the subset




Swiss bank notes: forward plot of minimum distances of units not in the subset, left-hand panel, and a zoom taken in the last 30 steps. There seems to be a cluster of outliers and a few more

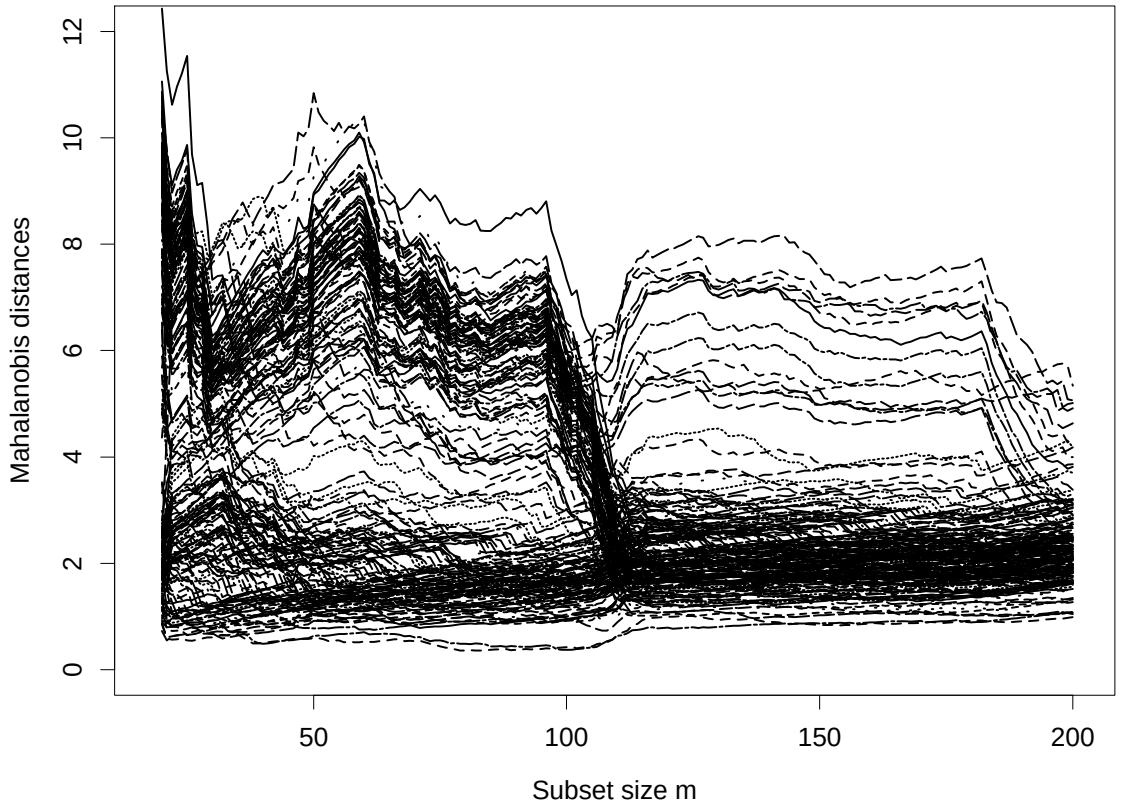
# Simulation Envelopes 3




- Swiss banknotes: forward plot of minimum distances of units not in the subset.
- The outliers are clear, but how many are there?

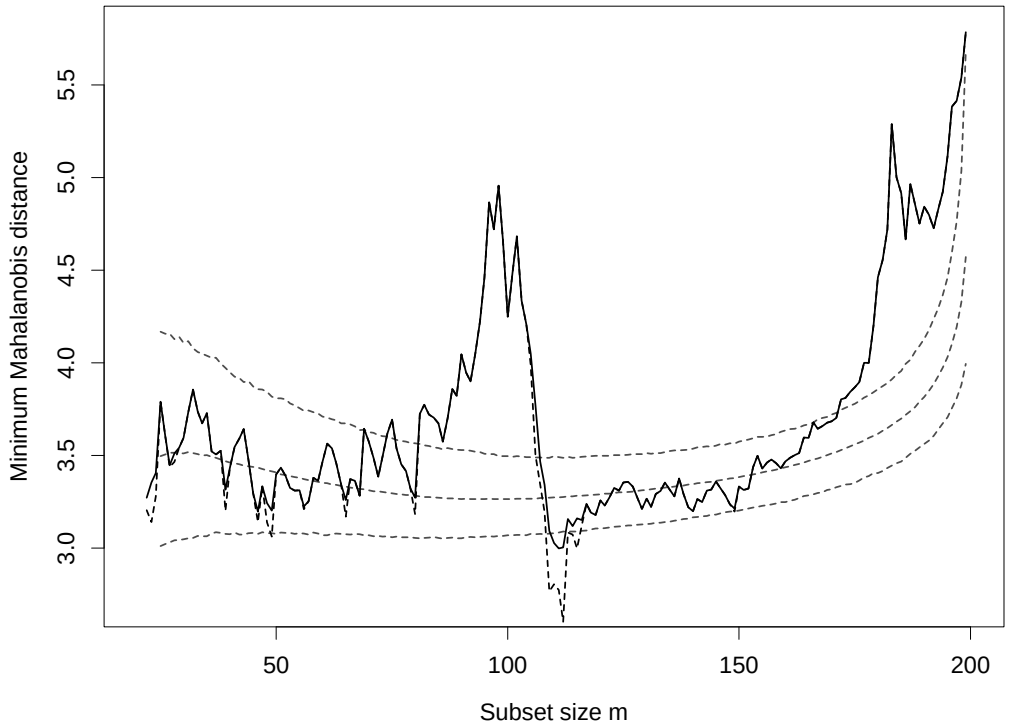
# Swiss Banknote Data 3

Now we start in the first group, that of supposedly genuine notes




Swiss bank notes, starting with the first 20 observations on genuine notes: forward plot of scaled Mahalanobis distances. Two groups are evident

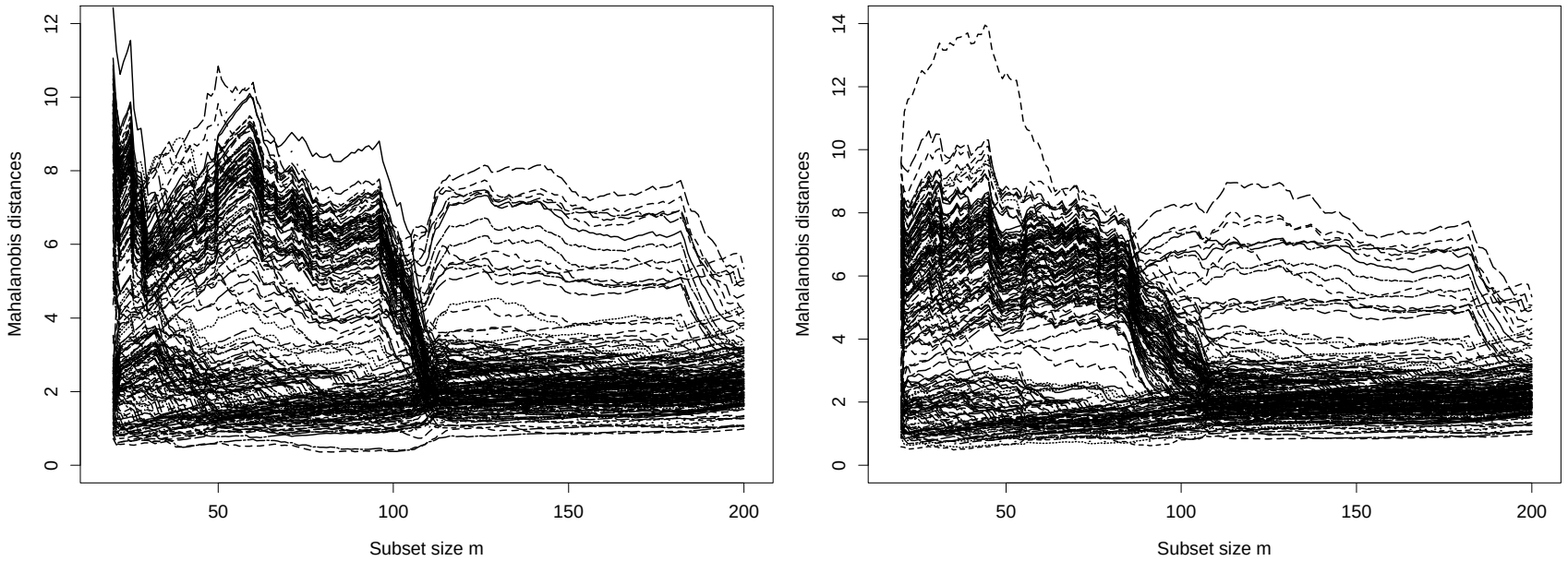
# Simulation Envelopes 4




- Swiss banknotes: forward plot of minimum distances of units not in the subset, starting with “genuine" notes.
- The two groups are now clear

# Swiss Banknote Data

- We started with a subset of “genuine" notes
- Now start with “forgeries"

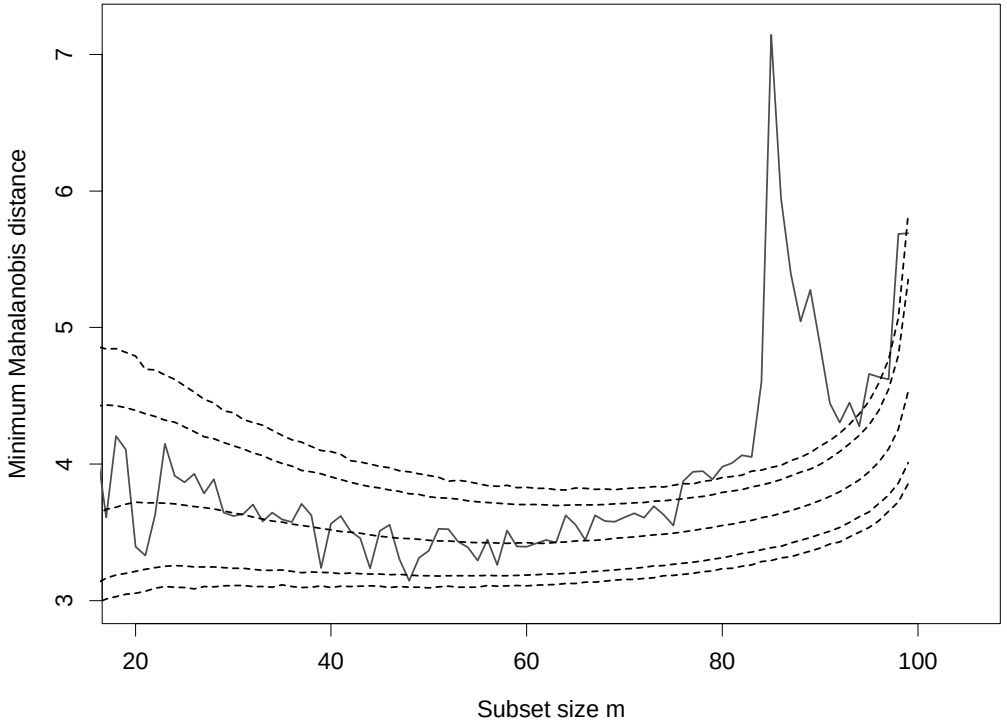



- Swiss bank notes: left panel, starting with the first 20 observations on genuine notes; right panel, starting with the first 20 observations on forgeries. Forward plots of scaled Mahalanobis distances
- The second halves of the two plots are similar

# Simulation Envelopes 5

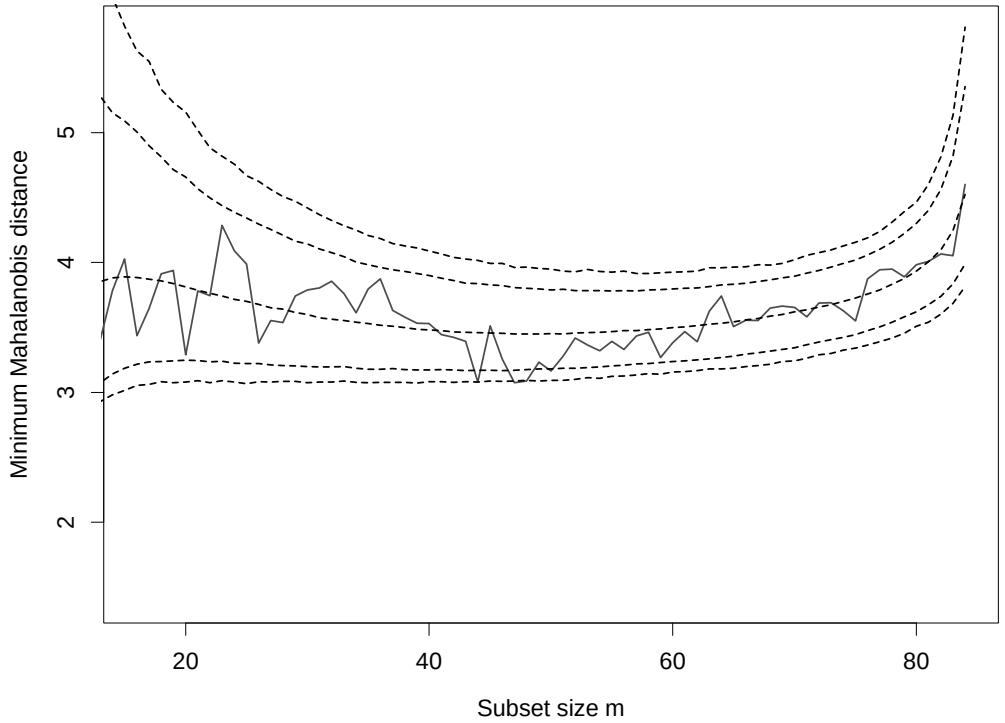
Swiss banknotes: it is also informative to look at plots of just the group of forgeries

# Simulation Envelopes 6




- Swiss banknotes - forgeries: forward plot of minimum distances of units not in the subset
- The outliers remain clear

# Simulation Envelopes 8

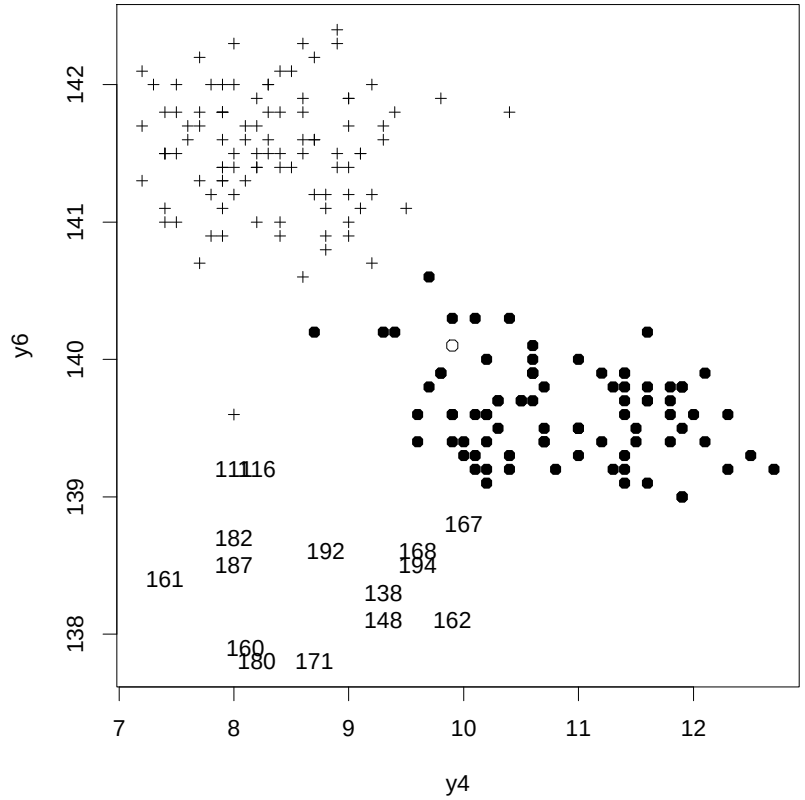



- Swiss banknotes - forgeries with 15 observations deleted: forward plot of minimum distances of units not in the subset.
- There is now no evidence of any further outliers

# Swiss Banknote Data 4

We can of course produce many other plots:

- Highlight individual Mahalanobis distances for each unit
- Forward plot of elements of covariance matrix
- Tentatively separate the groups and repeat the analyses
- . . .
- Plot the data




Swiss bank notes. Scatterplot of $y_6$ against $y_4$ which reveals most of the structure of all 200 observations: there are three groups and an outlier from Group 1, the crosses

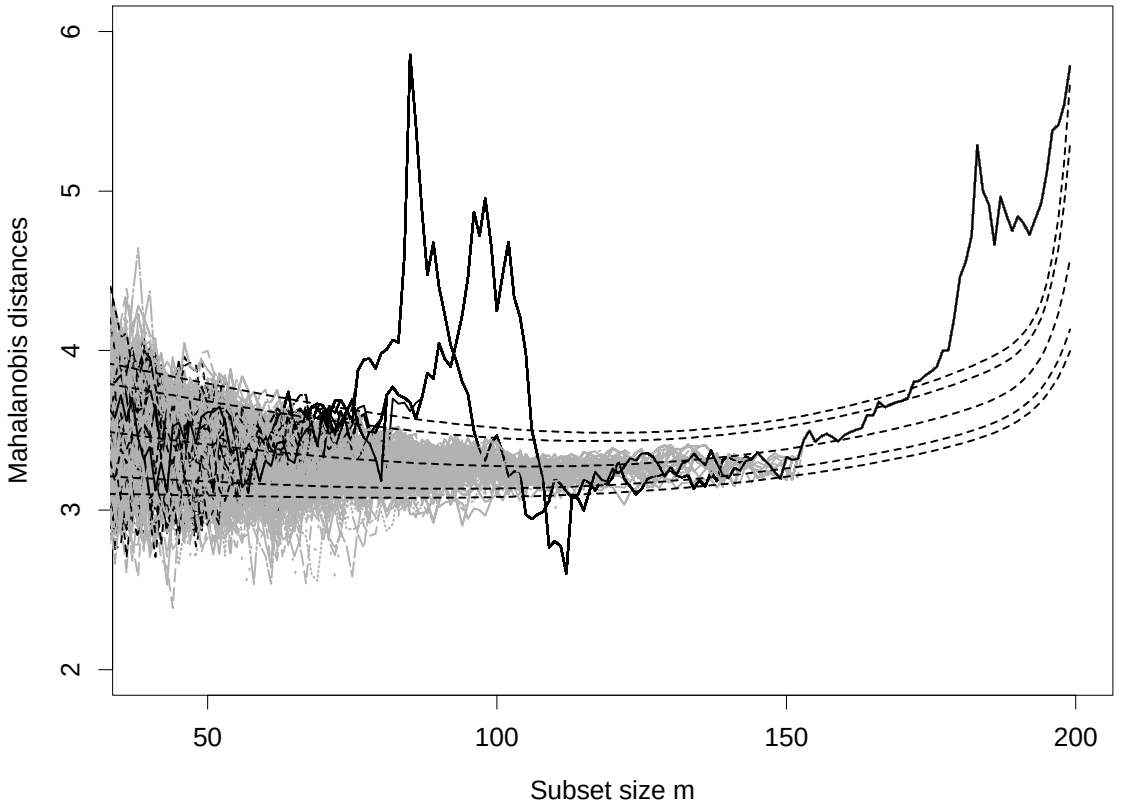
# Unknown Populations

- The preceding is not quite the point
- Starting from a plot of univariate distances may reveal structure
- But groups not usually known.
- Instead consider random starts

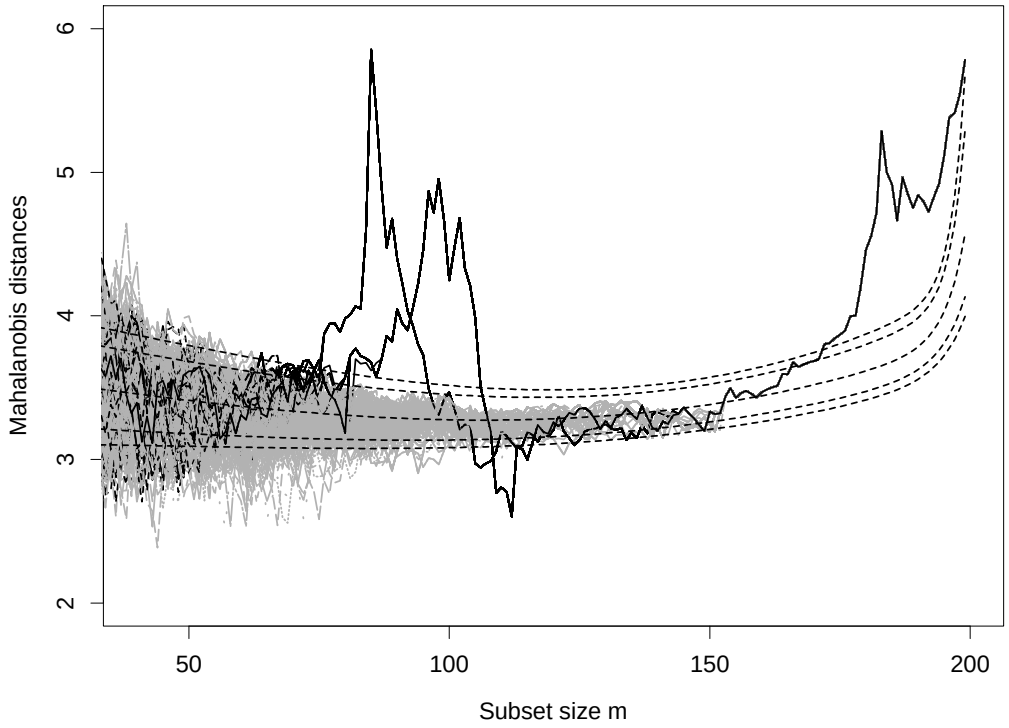
# Swiss Banknote Data 5

500 random starts with $m_0 = 10$.

- Also included in the plot are 1, 5, 50, 95 and 99% simulation envelopes for $d_{[m+1]}(m)$ when the observations come from a single six-dimensional normal distribution
- from $m$ around 150, all searches follow the same trajectory; the starting point is not of consequence in the latter part of the search




Swiss banknote data: minimum Mahalanobis distances amongst units not in the subset. 500 searches with random starting points; the searches shown in grey always contain units from both groups. The peaks contain 70 and 62 searches

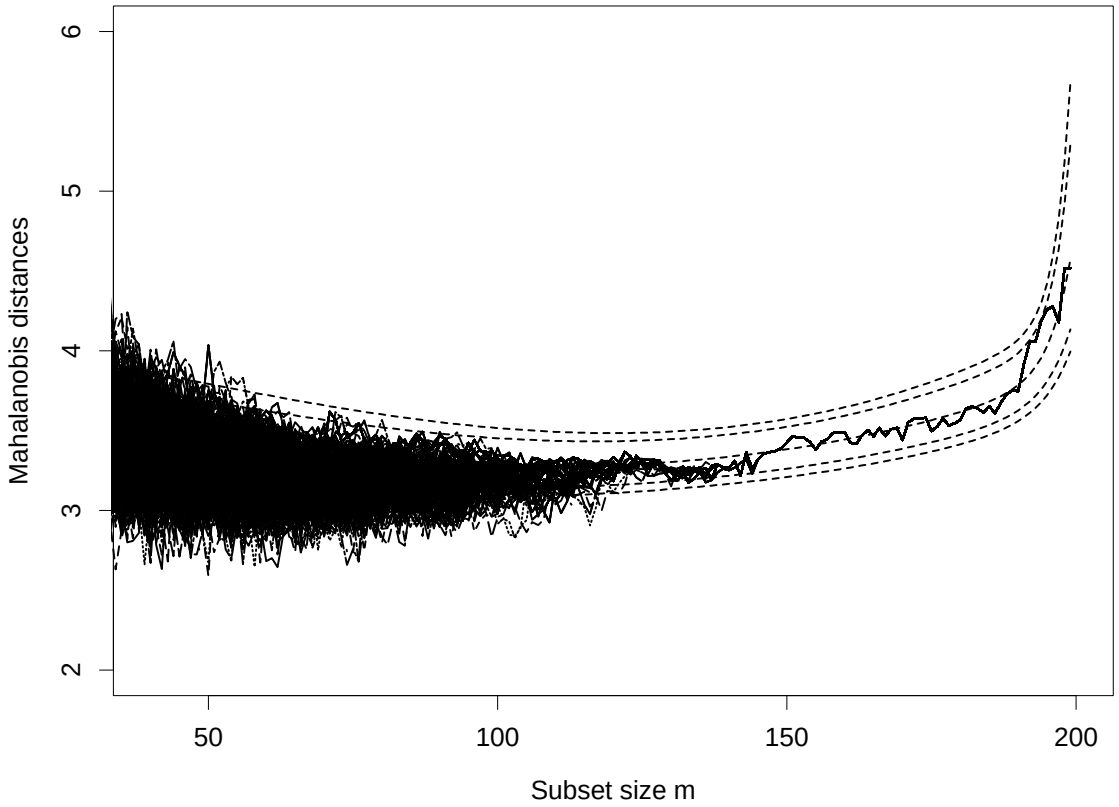



Swiss banknote data: 500 searches with random starting points

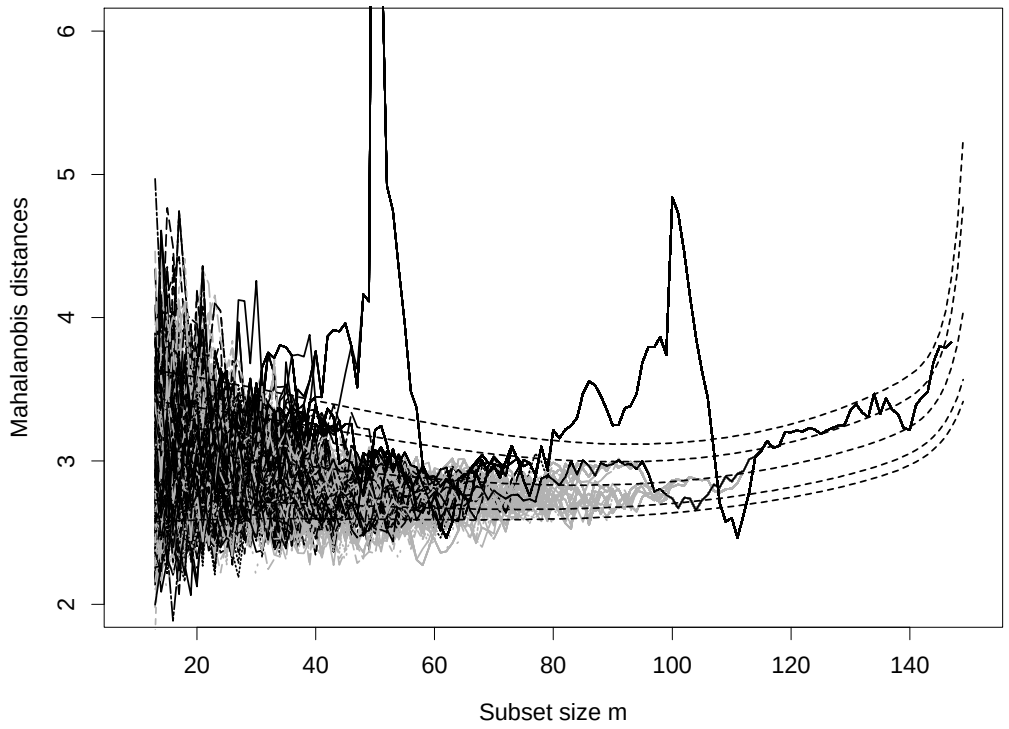
- The first peak (at $m = 85$) is from searches starting with the forgeries
- The second peak (around $m = 100$) is from searches starting with the genuine notes
- Searches are attracted towards individual groups

# Swiss Heads 4

We also need to demonstrate that we are not finding structure where none exists. The forward plot of the minimum distance of observations not in the subset shows none of the structure of clustering found in the banknote data. It however does show again how the search settles down in the last one third, regardless of starting point.



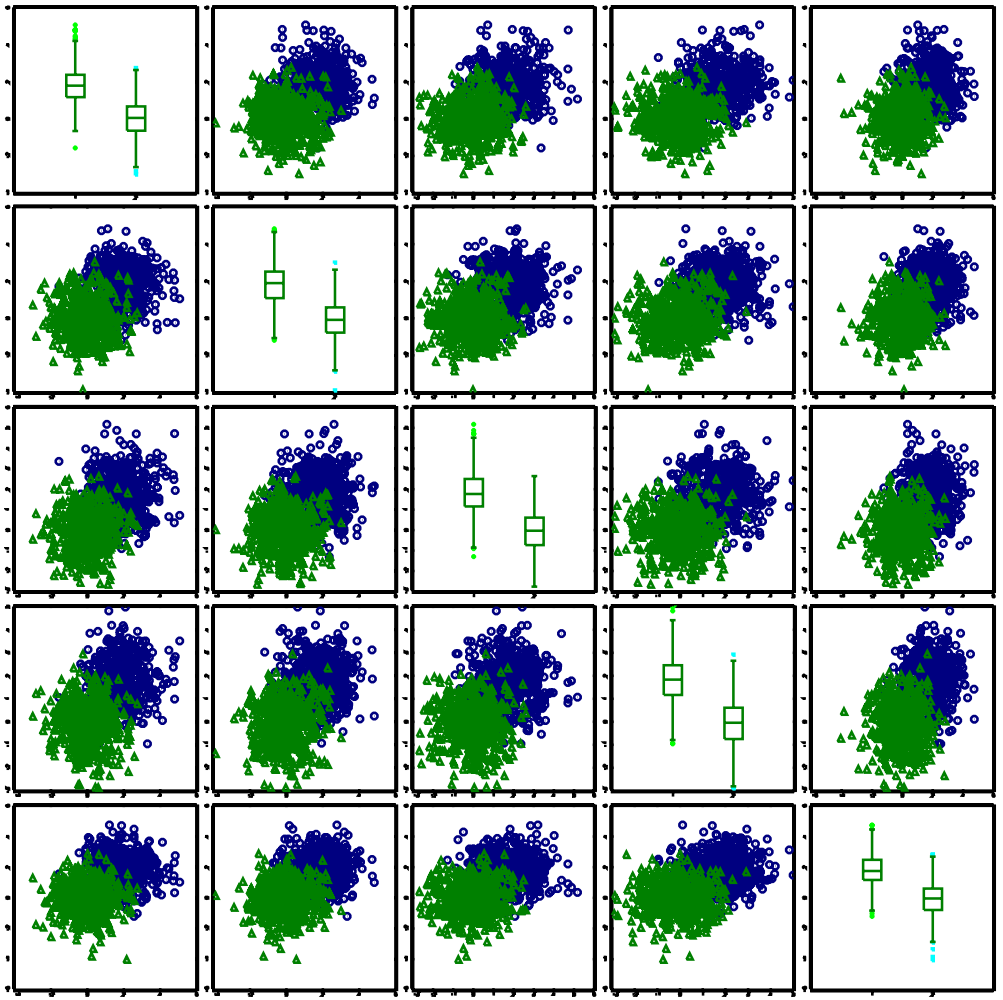

Swiss heads: minimum Mahalanobis distances amongst units not in the subset. 500 searches with random starting points. No evidence of any groups




Fisher's iris data: minimum Mahalanobis distances amongst units not in the subset. 500 searches with random starting points. The three well-known groups

# Interrogating the Plots

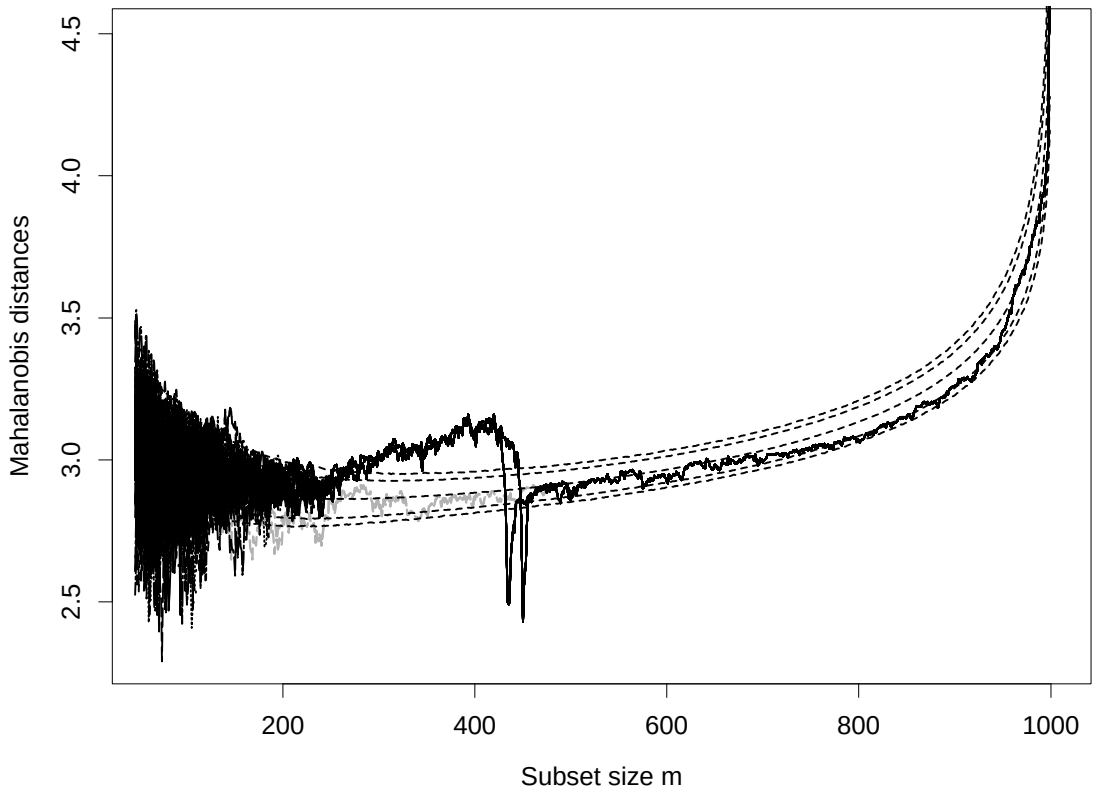
- The results of random starts provide a powerful visual method of detecting groups in data
- But how to quantify the groups?
- Look at a sample of 1,000 five-dimensional normal observations
- The samples have the same independent error structure, but differ slightly in mean
- There is appreciable overlapping.



Two clusters of independent normal variables: scatter plot matrix

# Interrogating the Plots 2

We look at 200 random start forward searches to elucidate the structure of the data.



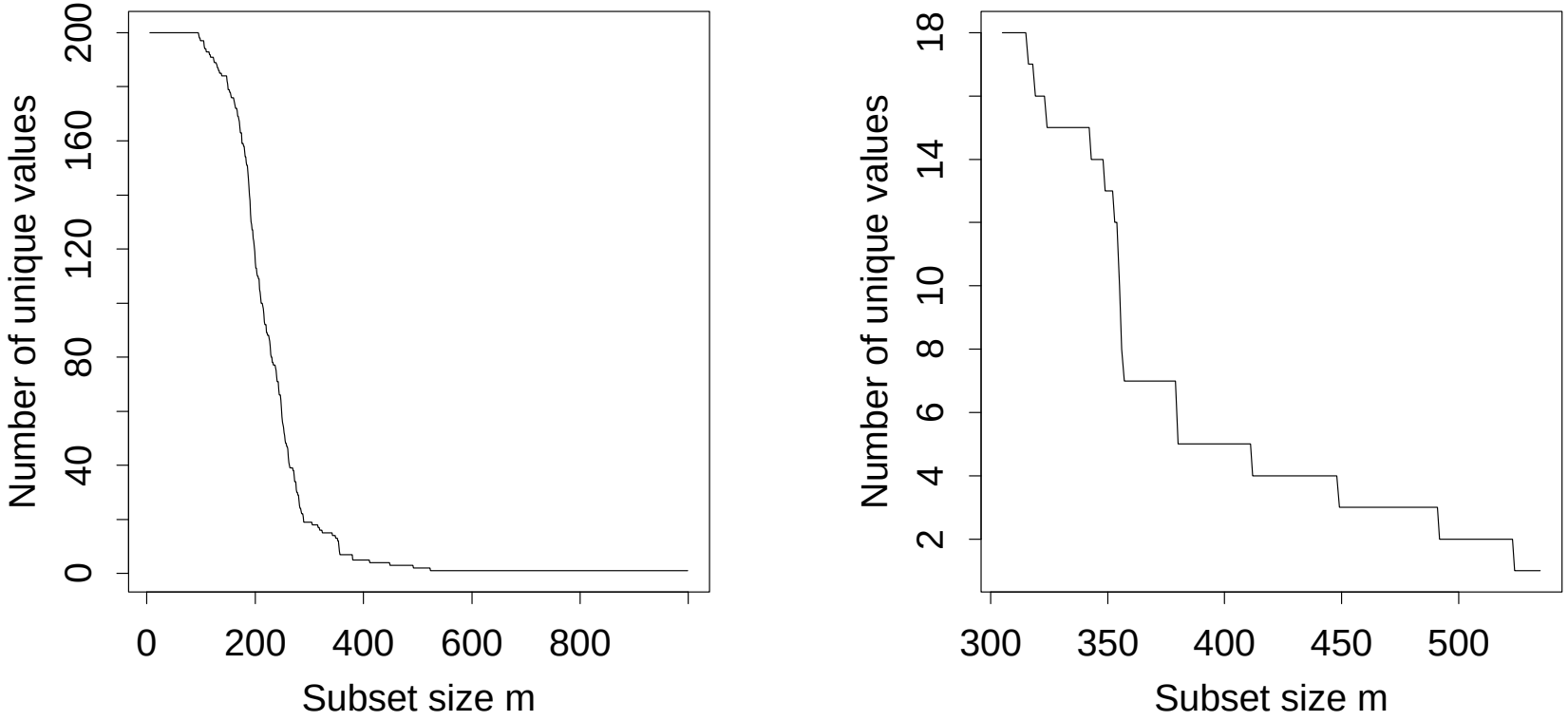

Two clusters of independent normal variables: forward plot of minimum Mahalanobis distances from 200 random starts. Two clusters are evident around $m = 400$. Trajectories in grey always include units from both groups

# Interrogating the Plots 3

- With 500 observations in each group we would like the dips below the envelopes to occur near $m = 500$
- However, around $m = 440$ both curves suddenly dip below the envelopes as relatively remote observations from the other group enter the subset
- The estimated covariance matrix is then slightly too large and the plotted distances are slightly small

# Interrogating the Plots 4

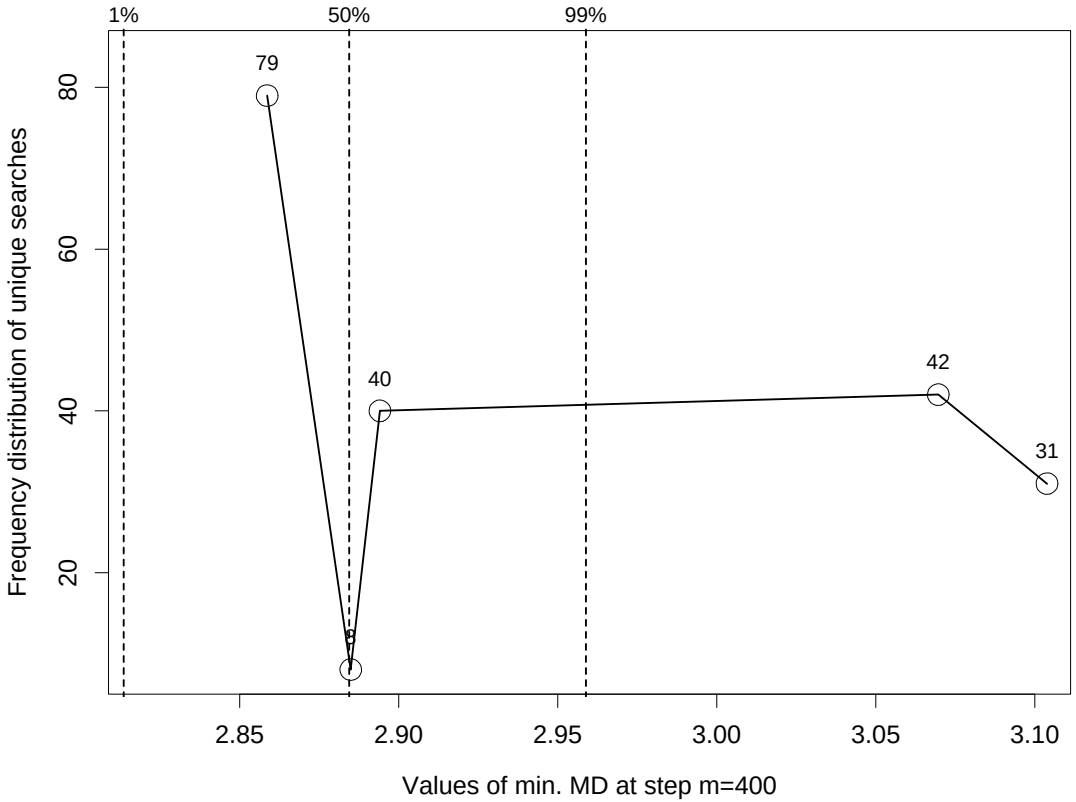
- Now consider cluster membership
- Initially there are 200 different values of $d_{\min}(m)$
- In the second half of the search, all 200 searches have converged and there is one value of $d_{\min}(m)$
- See how the number of different values of $d_{\min}(m)$ decreases with $m$




Two clusters of independent normal variables: forward plots of number of unique minimum Mahalanobis distances from 200 random starts. Left-hand panel, from 200 to 1; right-hand panel zoom of plot where clusters become apparent; there are five distinct values at $m = 400$

# Interrogating the Plots 5

- Now consider cluster membership
- Initially there are 200 different values of $d_{\min}(m)$
- In the second half of the search, all 200 searches have converged and there is one value of $d_{\min}(m)$
- To find the clusters we interrogate the figure at $m = 400$
- Find the subsets giving rise to the larger values of $d_{\min}(m)$

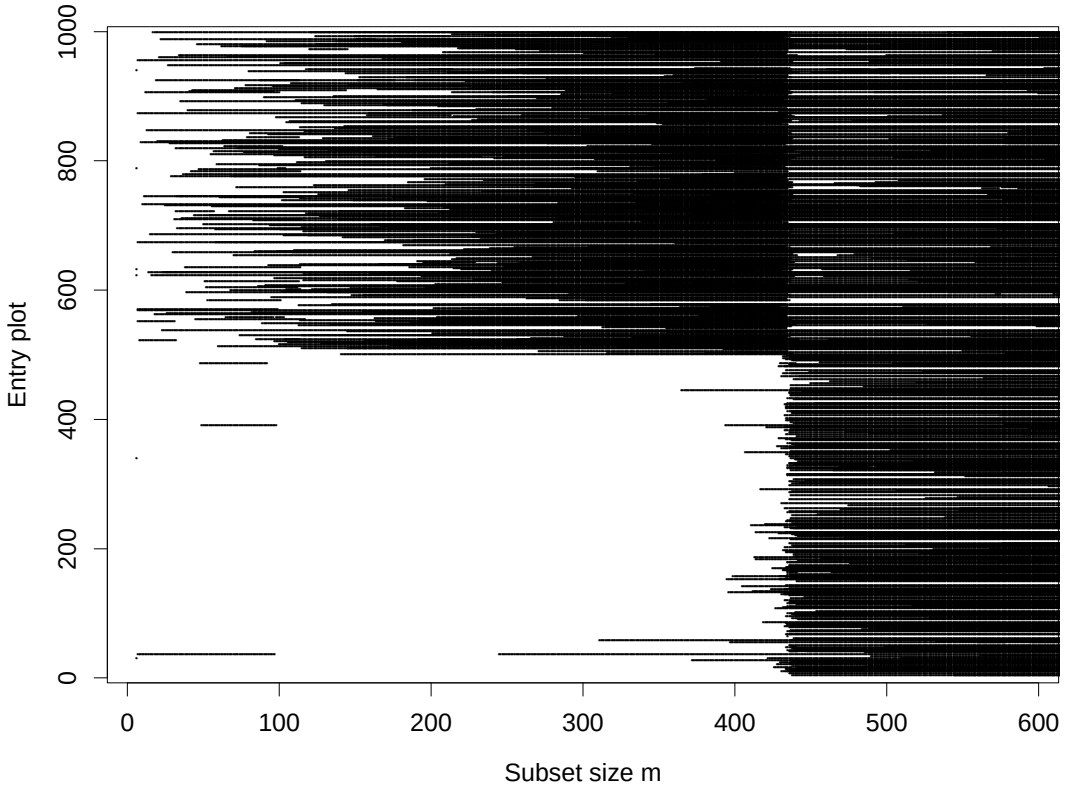



Frequency distribution of $d_{\min}(400)$ from 200 random starts.

- The vertical lines are the 1%, 50% and 99% points at $m = 400$ of the envelope
- There are two significantly large values indicating clusters

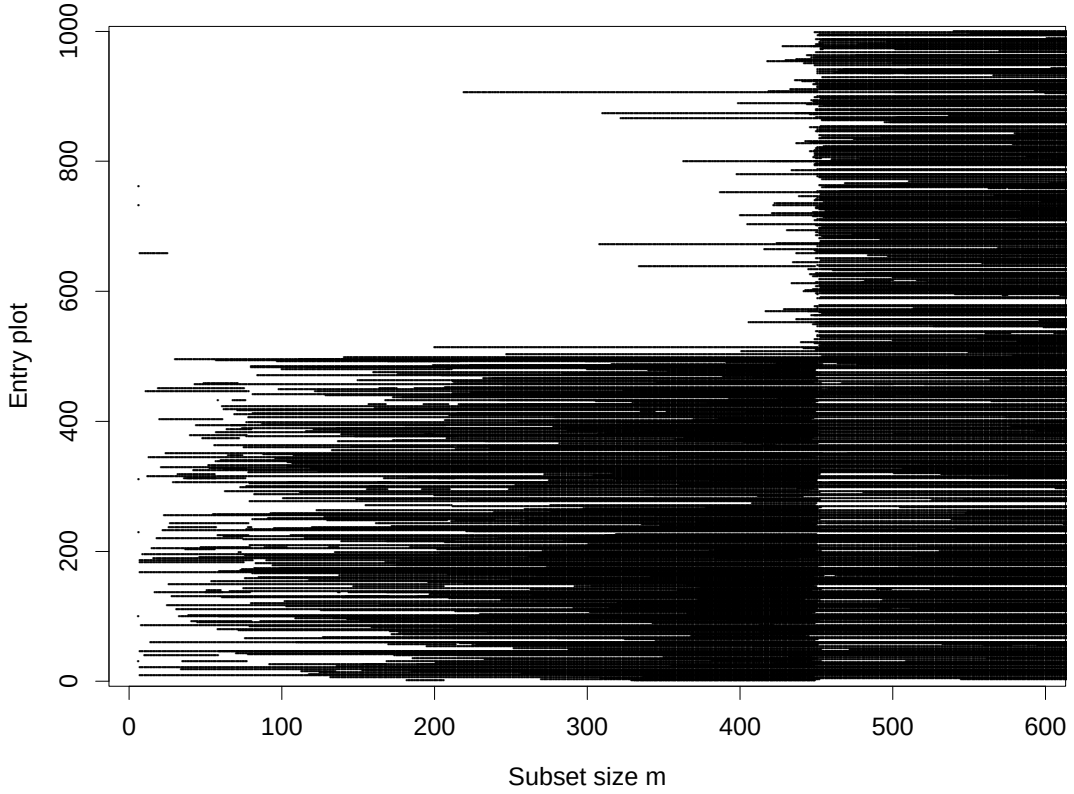
# Interrogating the Plots 6

- 31 searches give the most extreme value of $d_{\min}(m)$
- Since the searches have converged at $m = 400$, all will have the same residual trajectory
- We choose one at random to obtain a typical entry plot



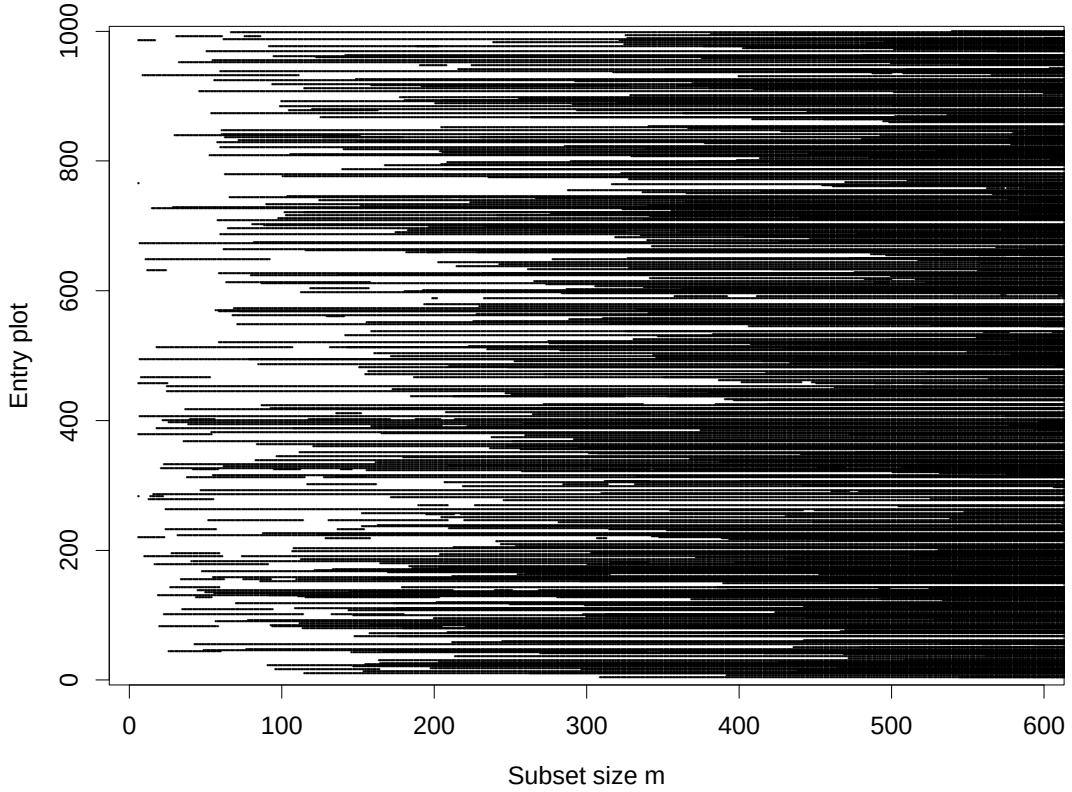

Two clusters of independent normal variables: entry plot for a trajectory yielding the highest value of $d_{\min}(400)$

- We have found the trajectories that include observations from those numbered 501 - 1,000
- Of course, labels not known for clustering




Two clusters of independent normal variables: entry plot for a trajectory yielding the second highest value of $d_{\min}(400)$

- Now we have found the trajectories that include observations from those numbered 1 - 500




Two clusters of independent normal variables: entry plot for a trajectory yielding the third highest value of $d_{\min}(400)$

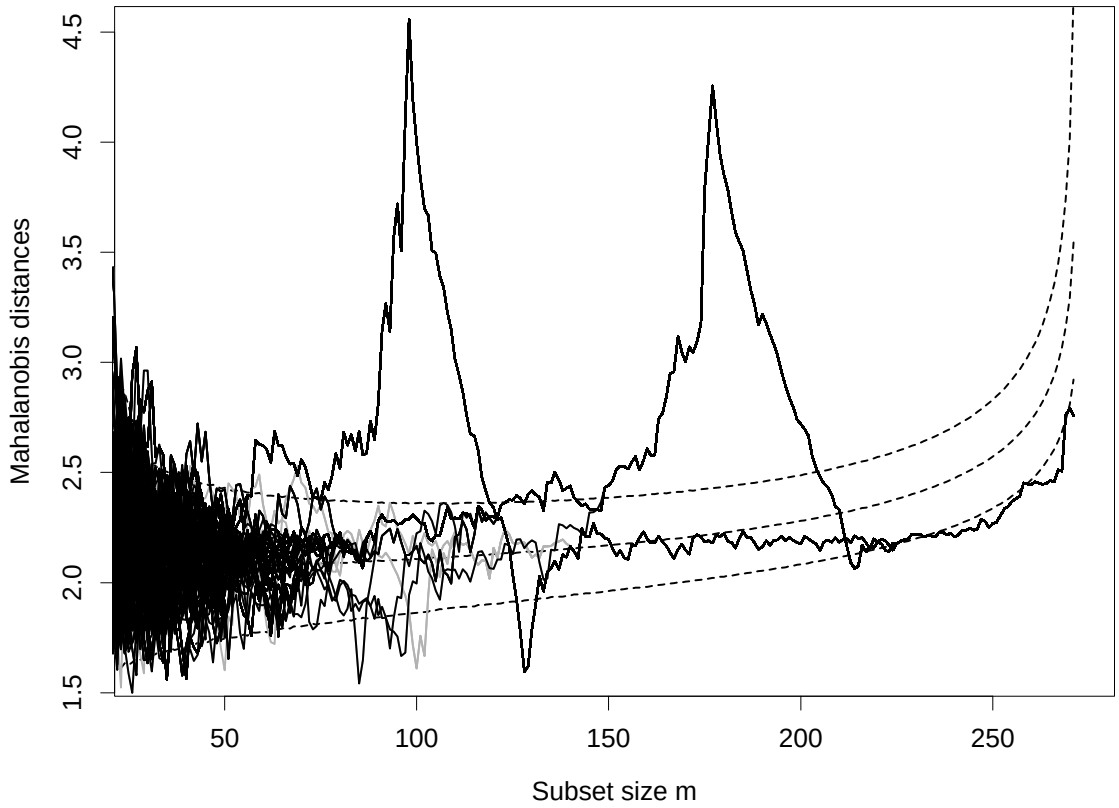
- Now we have a mixture of the two clusters

# Interrogating the Plots 7

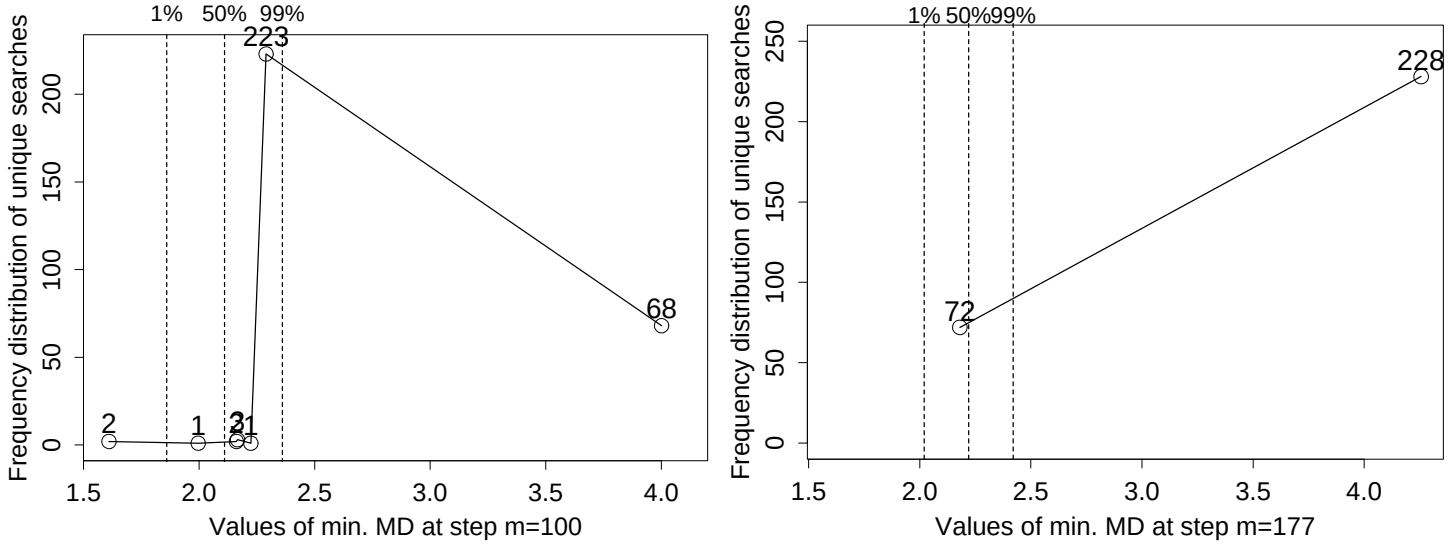
- The two clusters have been found
- The plots show the overlapping nature of the clusters as units from both are in the subset well before $m = 500$.
- Explore structure by a forward search starting from units in each tentative cluster in turn
- Confirm by search fitting two populations at once

# Comparison 1

- Eruptions of ‘Old Faithful’ from the MASS library
- 272 observations with $x_{1i}$ the duration of the $i$th eruption and $x_{2i}$ the waiting time to the start of that eruption from the start of eruption $i - 1$
- Here we use clustering to establish whether one, or more than one, mechanism is present and, if so, how many
- See (Azzalini and Bowman 1990) for a further discussion




Old Faithful data: forward plot of minimum Mahalanobis distances from 300 random starts with 1%, 50% and 99% envelopes. Two clusters are evident around $m = 99$ and $m = 178$. The few trajectories in grey always include units from both groups

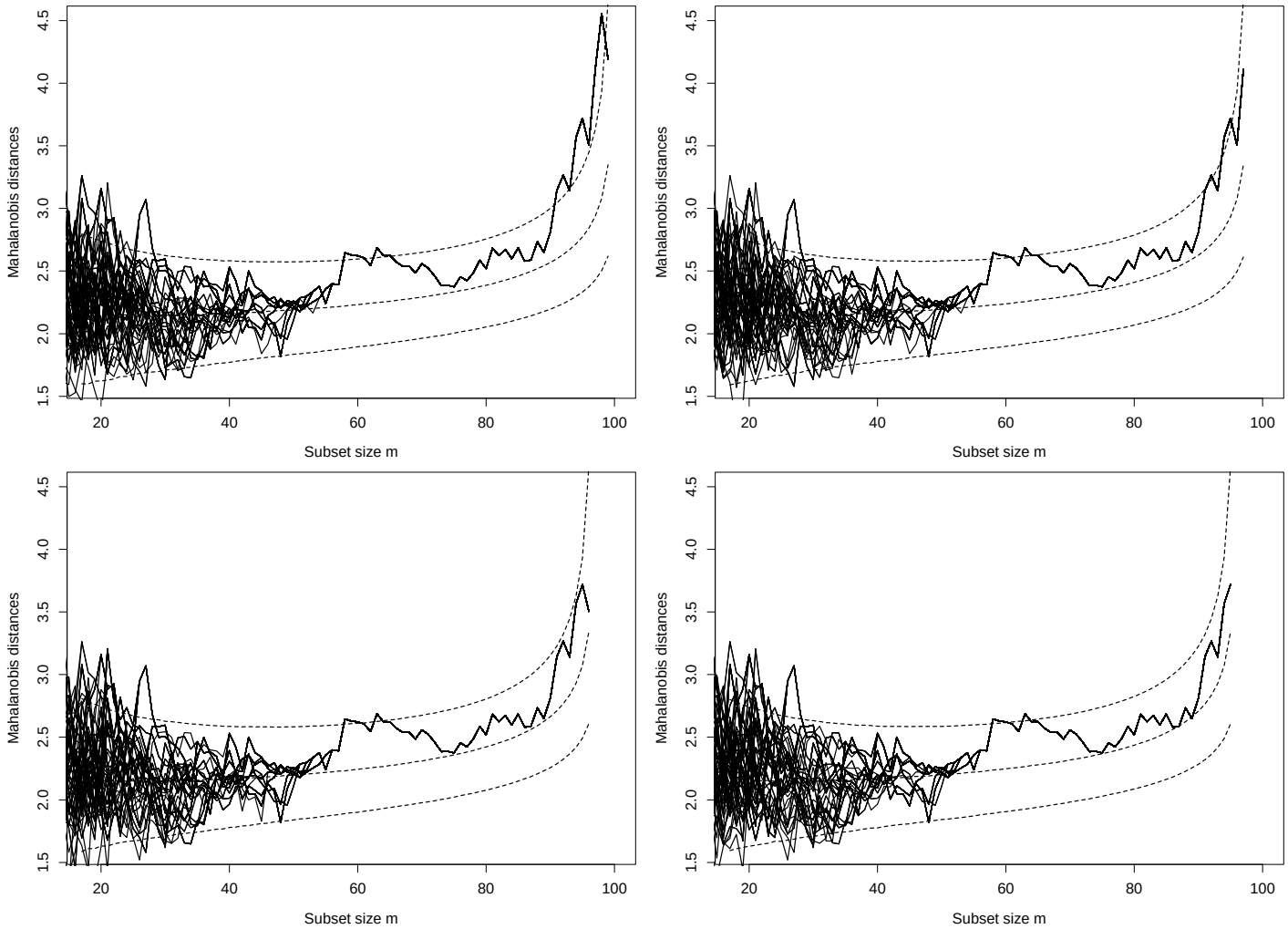



Old Faithful data: frequency distributions of (a) $d_{\min}(100)$ and (b) $d_{\min}(177)$. The vertical lines are the 1%, 50% and 99% points at $m$ of the envelope

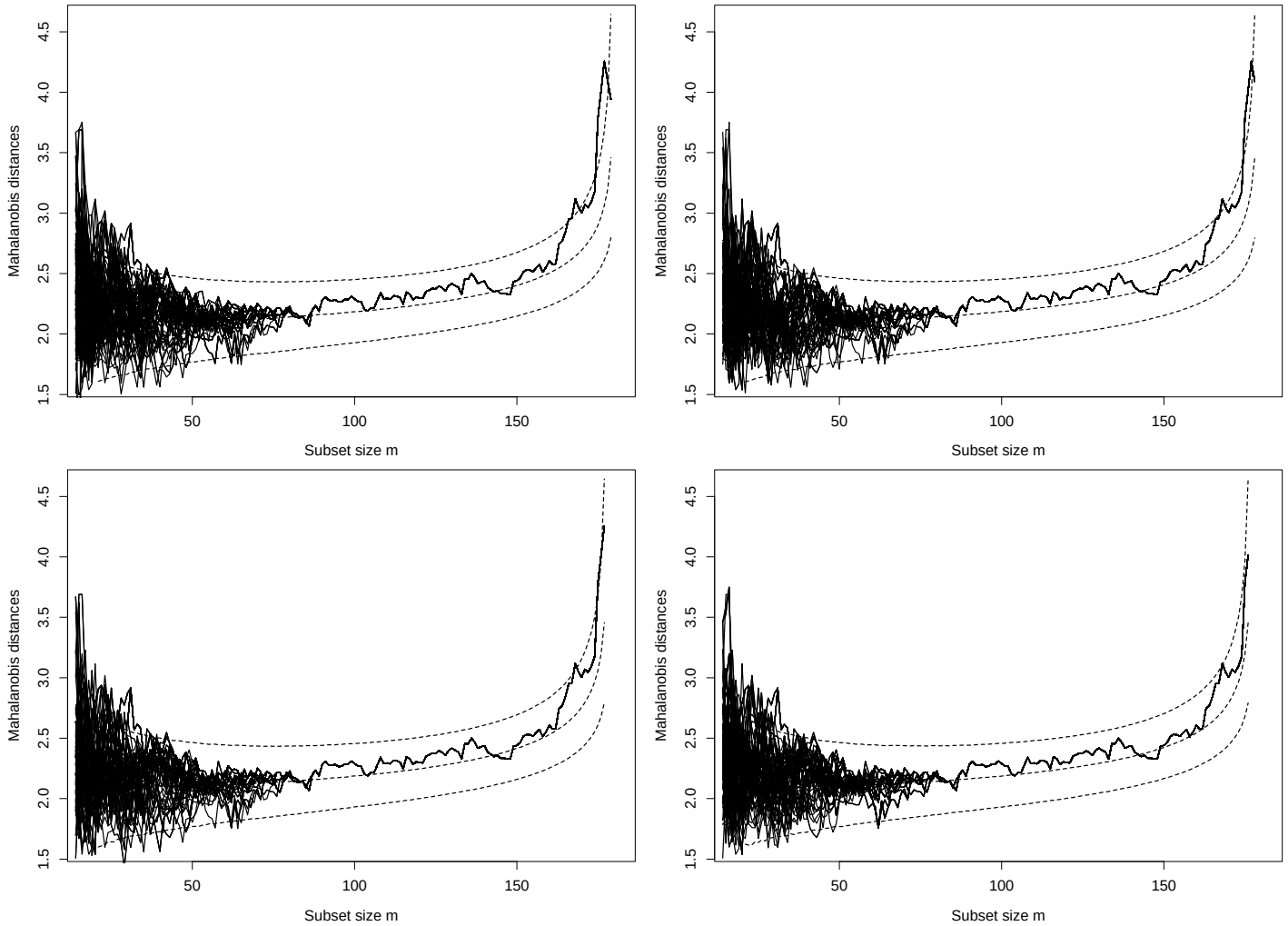
- The membership of the two clusters can now be established

# Comparison 2

- We now start forward searches from the two indicated clusters
- The exact cluster membership is obtained by testing against simulation bounds for various $n$



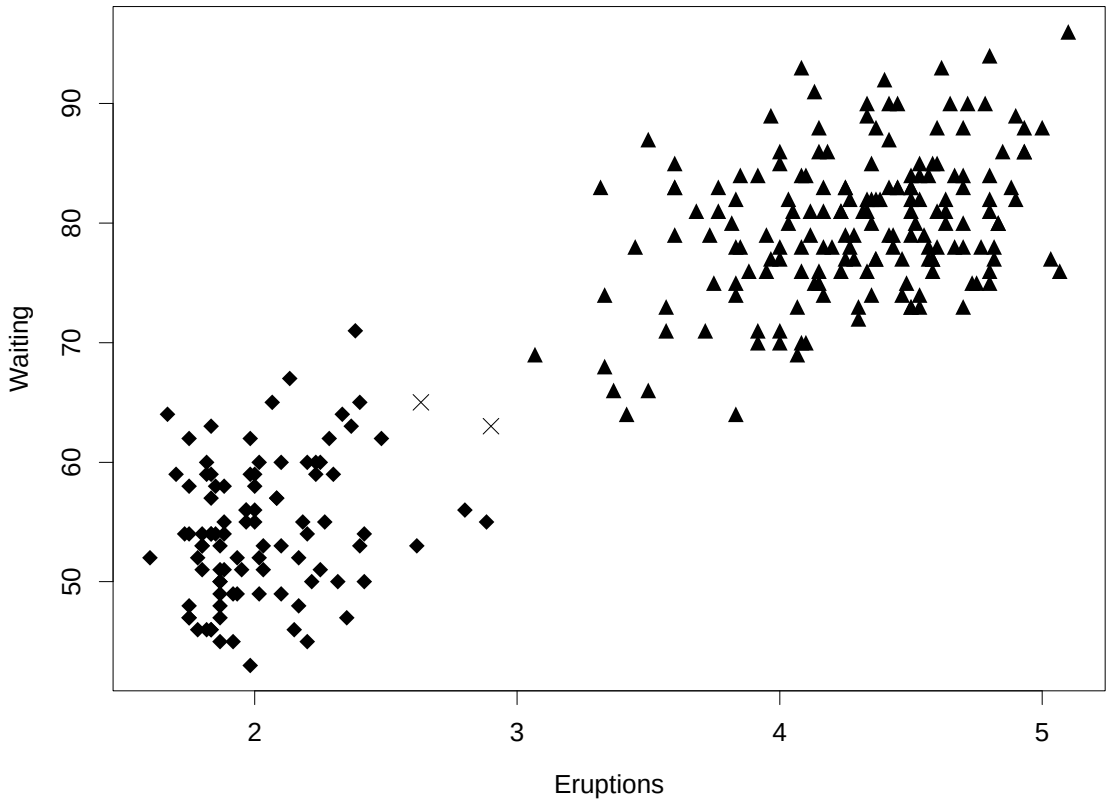

Old Faithful data: forward plots of minimum Mahalanobis distances from 100 random starts for the data forming the first cluster: (a) $n = 100$; (b) $n = 98$; (c) $n = 97$; (d) $n = 96$. A cluster of 97 units is indicated




Old Faithful data: forward plots of minimum Mahalanobis distances from 100 random starts for the data forming the second cluster: (a) $n = 180$; (b) $n = 179$; (c) $n = 178$; (d) $n = 177$. A cluster of 177 units is indicated

# Comparison 3

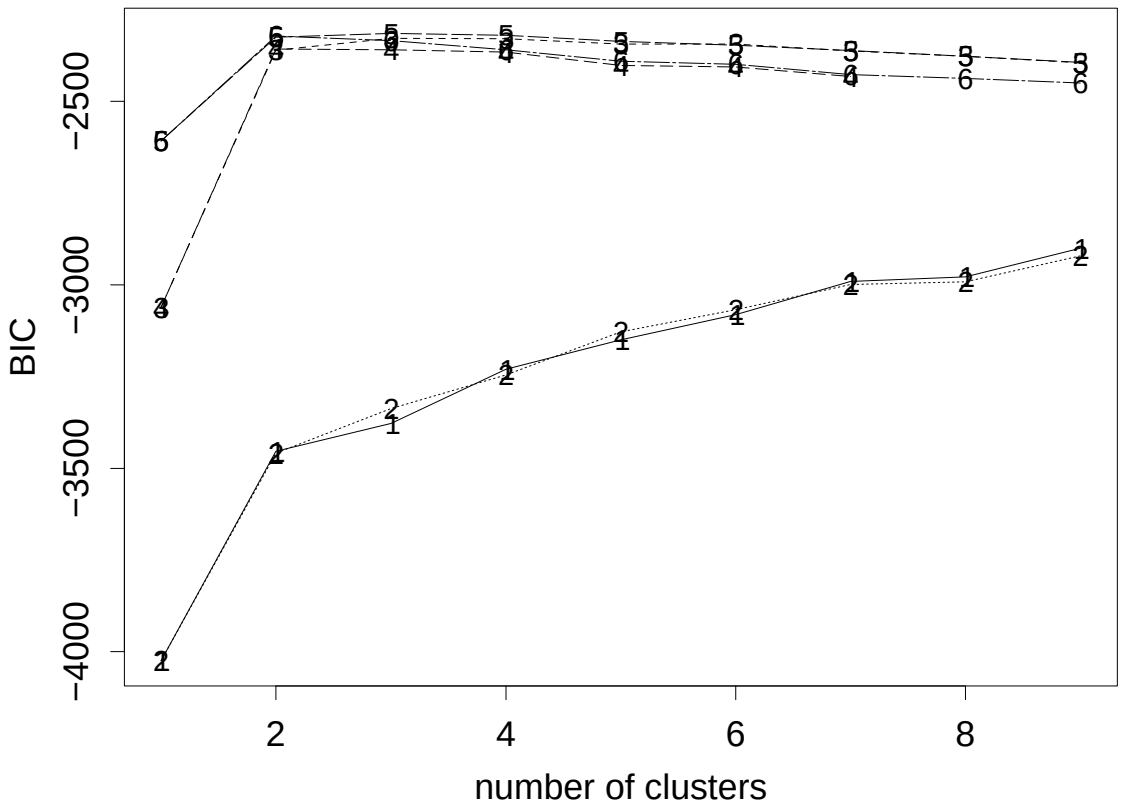
- We have two clusters, one of 97 units and the other of 177, making 274 units in all, two more units than there are in the data
- As the figure shows, there are no outliers and two units could belong to either cluster



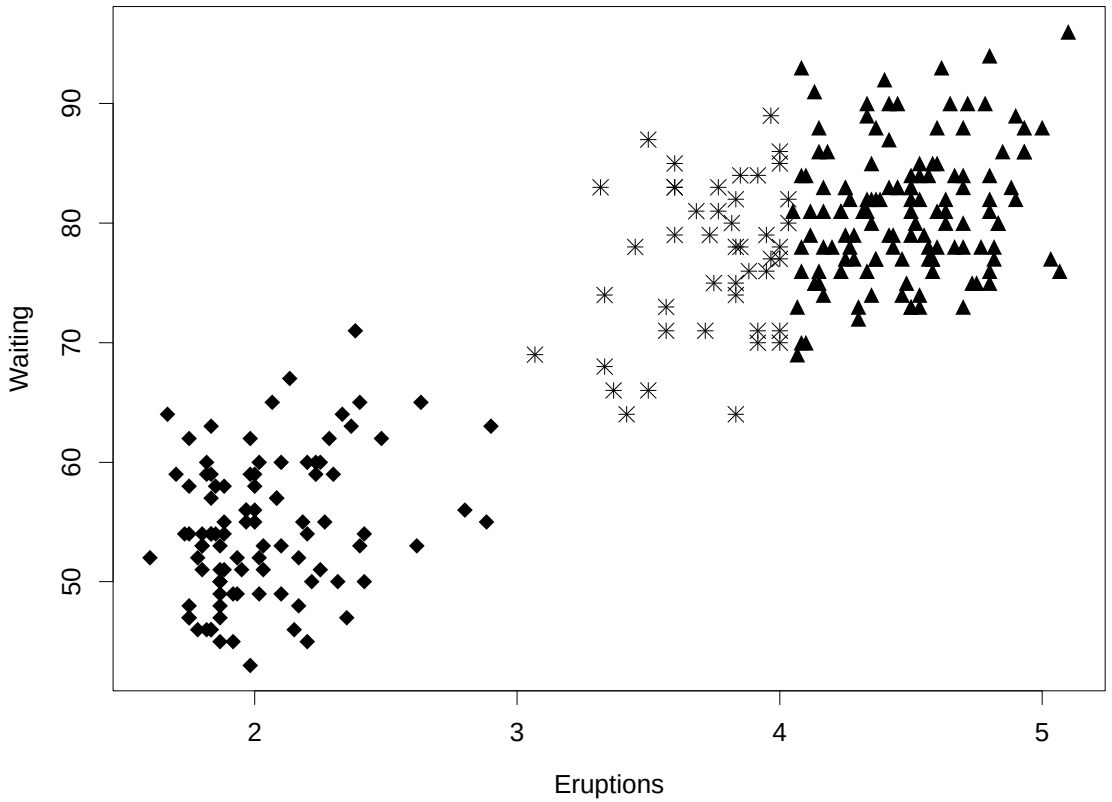

Old Faithful data: scatterplot matrix showing the two clusters. The units marked $\times$ could lie in either cluster (the distances are not in Euclidean space)

# Comparison 4

- Now compare with `mclust` which fits mixtures of normal distributions with a variety of covariance structures
- Model is chosen using BIC




Old Faithful data: BIC plot from `mclust`. The favoured model has three clusters with similarly shaped covariance matrices




Old Faithful data: scatterplot matrix showing the three clusters favoured by `mclust`. Too high a priority on similar covariance matrices?

# References

Azzalini, A. and A. Bowman (1990). A look at some data on the Old Faithful geyser. *Applied Statistics 39*, 357–365.