Regression Diagnostics and the Forward Search 3. A Single Multivariate Sample

Anthony Atkinson, LSE

Multivariate Normality

Much multivariate data is modelled with the normal distribution, often after a transformation to approximate normality (Box and Cox 1964). But do we have:

- A sample from a single normal population?
- The same, but with some outliers?
- A sample from several normal populations?
- The same with outliers as well?

The numbers of populations and of outliers are both unknown

Obscured Structure

The main diagnostic tools that we use are:

1. Plots of the data, especially scatterplot matrices
2. Various plots of Mahalanobis distances. The squared distances for the sample are defined as

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

   where $\hat{\mu}$ is the vector of means of the n observations and $\hat{\Sigma}$ is the unbiased estimator of the population covariance matrix.
3. These are the multivariate form of scaled residuals.

But: 1. Hard to interpret for many variables ; 2. Subject to masking.

The Forward Search 1

We use the Forward Search to find structure:

- Explore relationship between data and fitted models that may be obscured by fitting (masking)
- Output mostly graphical (versions of tests).
- FS orders the observations by closeness to the assumed model
- **Start** with a small subset of the data
- **Move Forward:** increase the number of observations m used for fitting the model.
- Continue until $m = n$

The Forward Search 2

For a subset of m observations the parameter estimates are $\hat{\mu}(m)$ and $\hat{\Sigma}(m)$. From this subset we obtain n squared Mahalanobis distances

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

- When m observations are used in fitting, the optimum subset $S^*(m)$ yields n squared distances $d_i^2(m^*)$
- Order these squared distances and take the observations corresponding to the $m+1$ smallest as the new subset $S^*(m+1)$
- Usually this process augments the subset by only one observation. Sometimes two or more observations enter as one or more leave

The Forward Search 3: One Population

- For each $m_0 \leq m \leq n$, plot the n distances $d_i(m^*)$, a forward plot.
- The starting subset of m_0 ($< n/10$) comes from bivariate boxplots that exclude outlying observations in any one or two-dimensional plot
- Content of contours adjusted to give required m_0
- **With one population** the search is not sensitive to the exact choice of starting subset.

The Forward Search 4

The distances tend to decrease as n increases. If interest is in the latter part of the search we look at

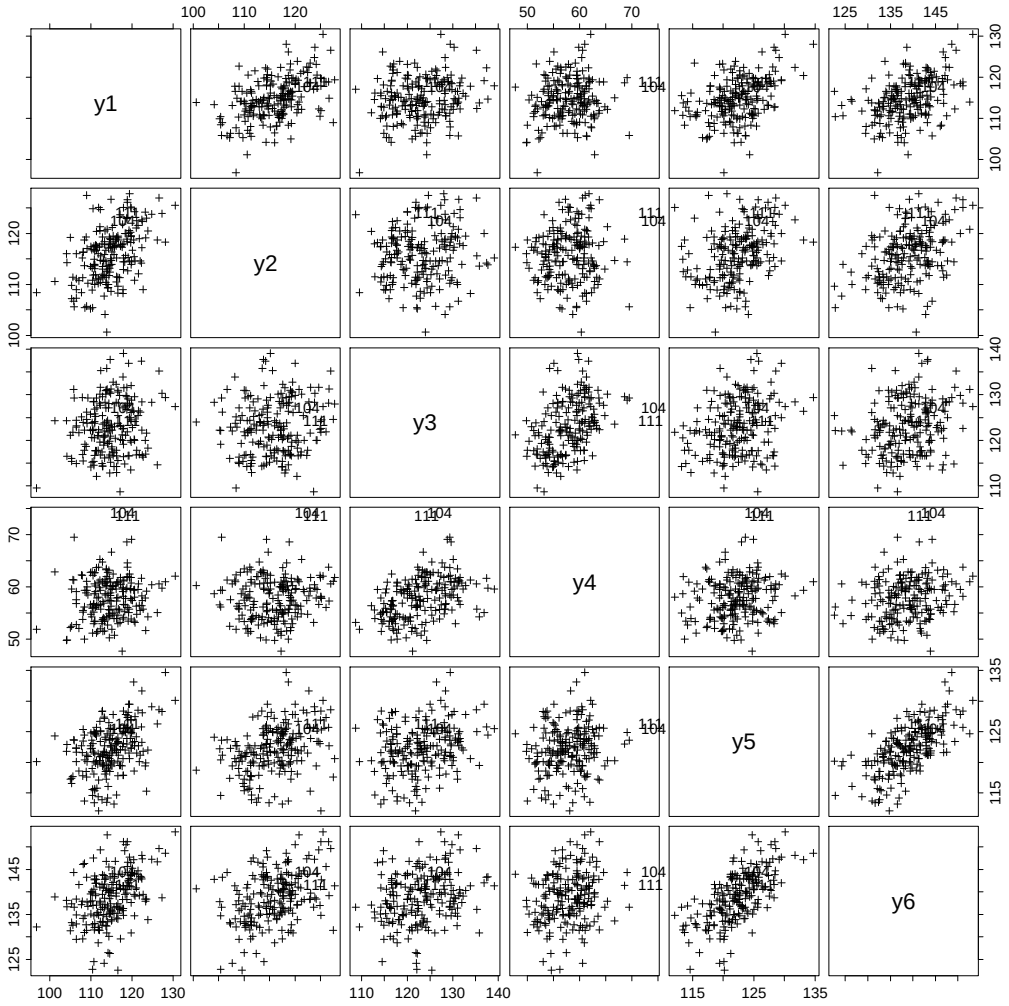
- **Scaled** distances

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

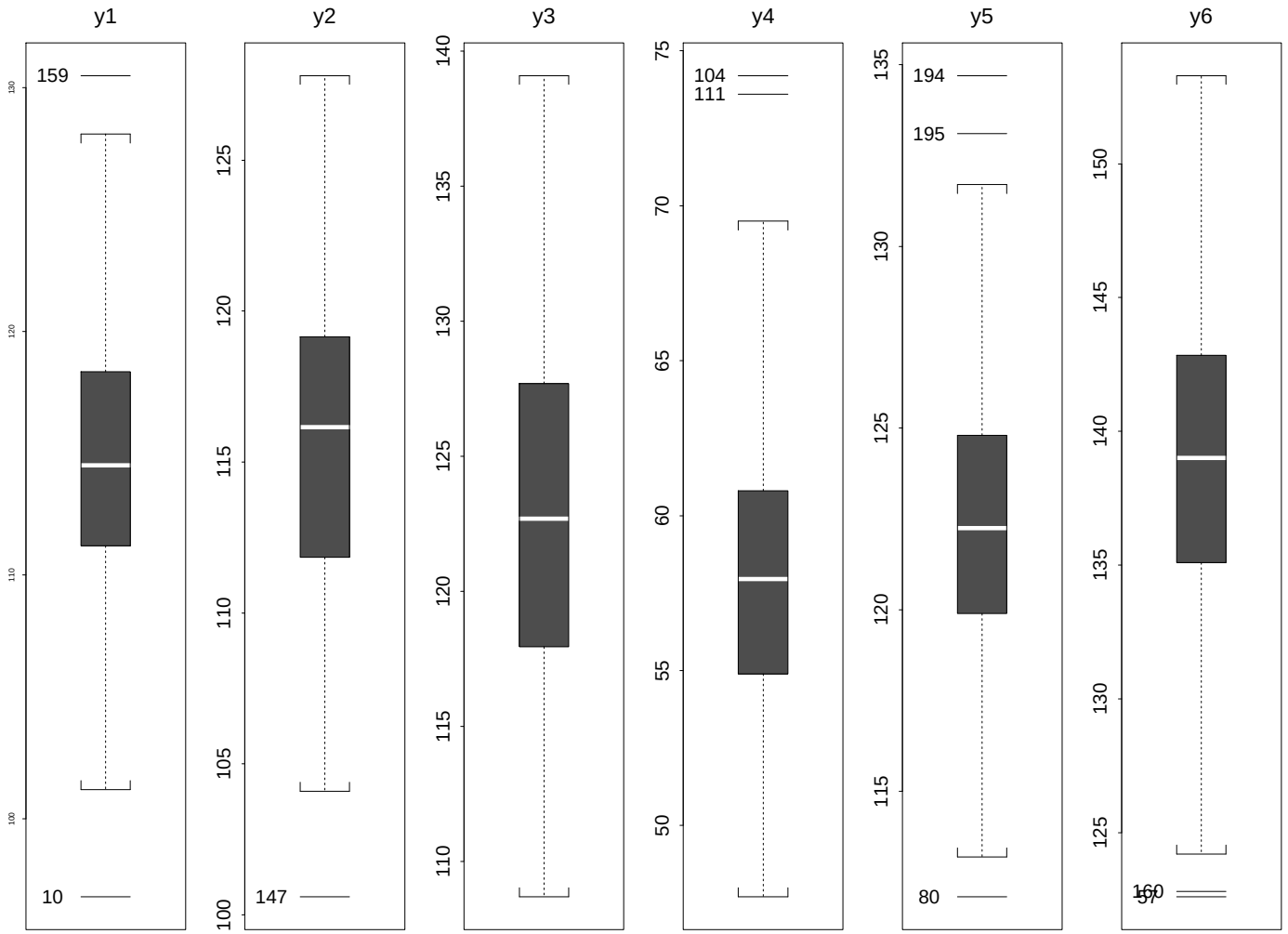
- v is the dimension of the observations y (v variables) and $\hat{\Sigma}(n)$ is the estimate of Σ at the end of the search.

Swiss Heads 1

As a first example of the use of forward plots we start with data given by Flury and Riedwyl (1988, p. 218): six readings on the dimensions of the heads of 200 twenty year old Swiss soldiers.



Swiss heads: scatterplot matrix with observations 104 and 111 marked



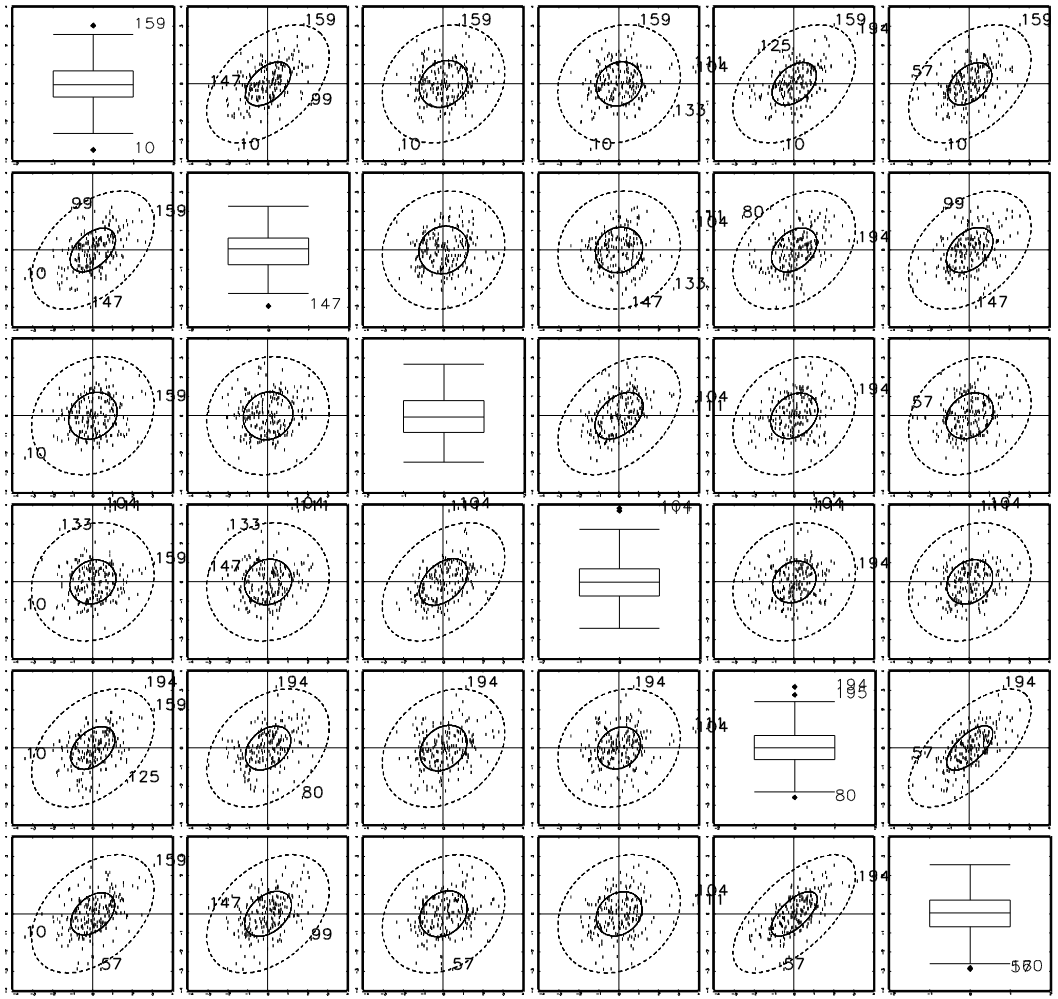
Swiss heads: boxplots of the six variables with univariate outliers labelled

Swiss Heads

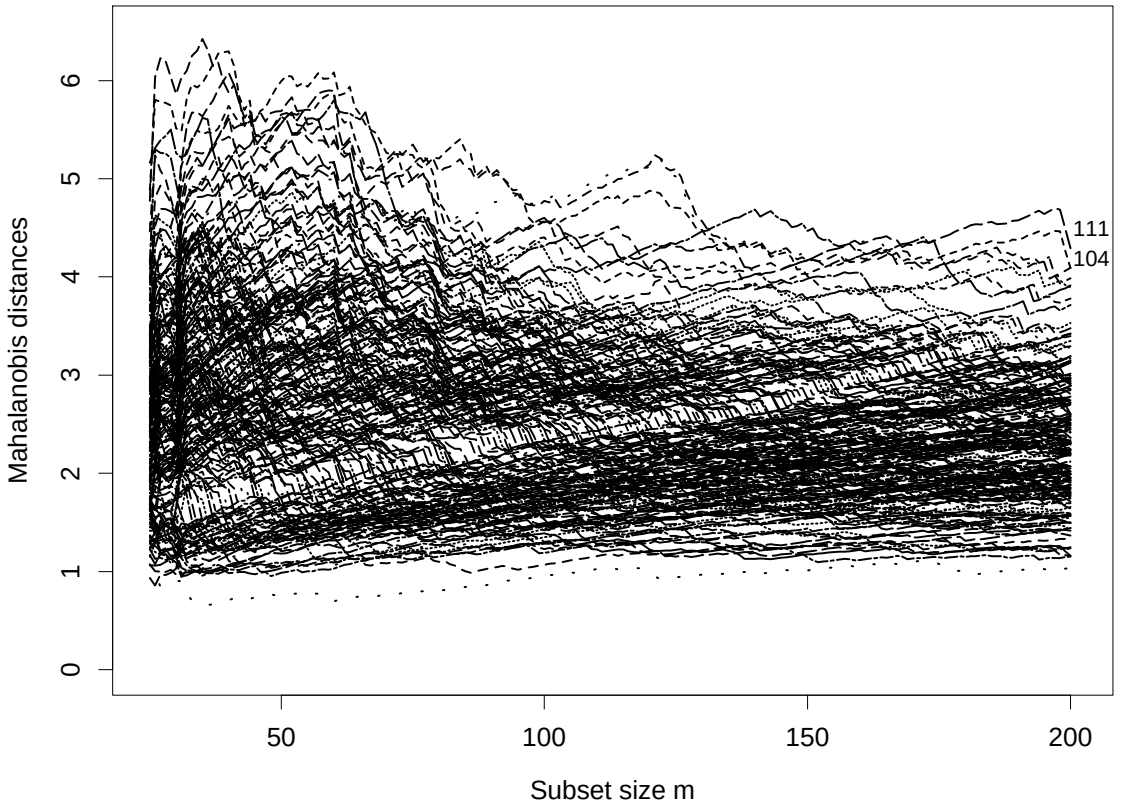
Starting the Search. We find observations within elliptical contours fitted to all the data.

The scaling parameter for the ellipses is called θ, the value being chosen to give the desired value for m_0.

The distribution of the $d_i^2(n)$ is scaled Beta, approximated by a scaled F distribution - exact if Σ estimated but μ known. The value of θ can be interpreted as a quantile of the F



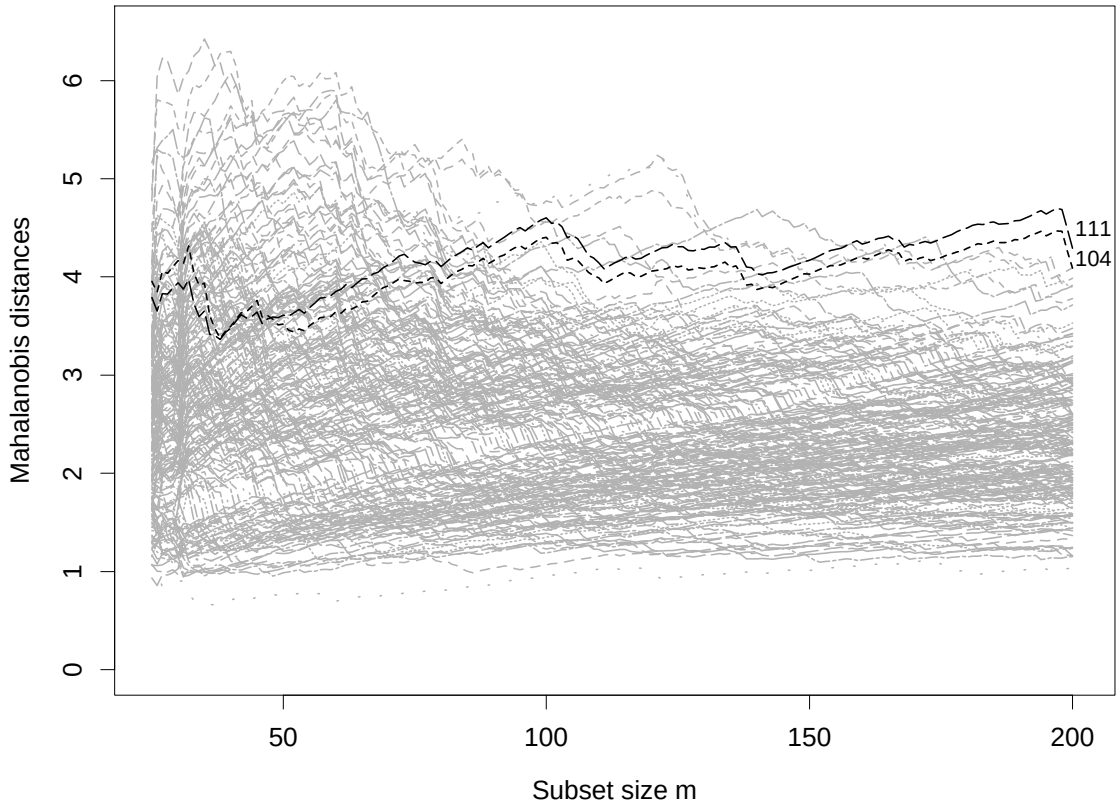
Swiss heads: scatterplot matrix. The outer ellipse ($\theta = 4.71$) indicates some potential outliers. The inner ellipse ($\theta = 0.92$) gives m_0 = 25 ???



Swiss heads: forward plot of scaled Mahalanobis distances showing little structure. The rising diagonal white band separates those units which are in the subset from those that are not. At the end of the search there are perhaps two outliers, observations 104 and 111.

Swiss Heads 2

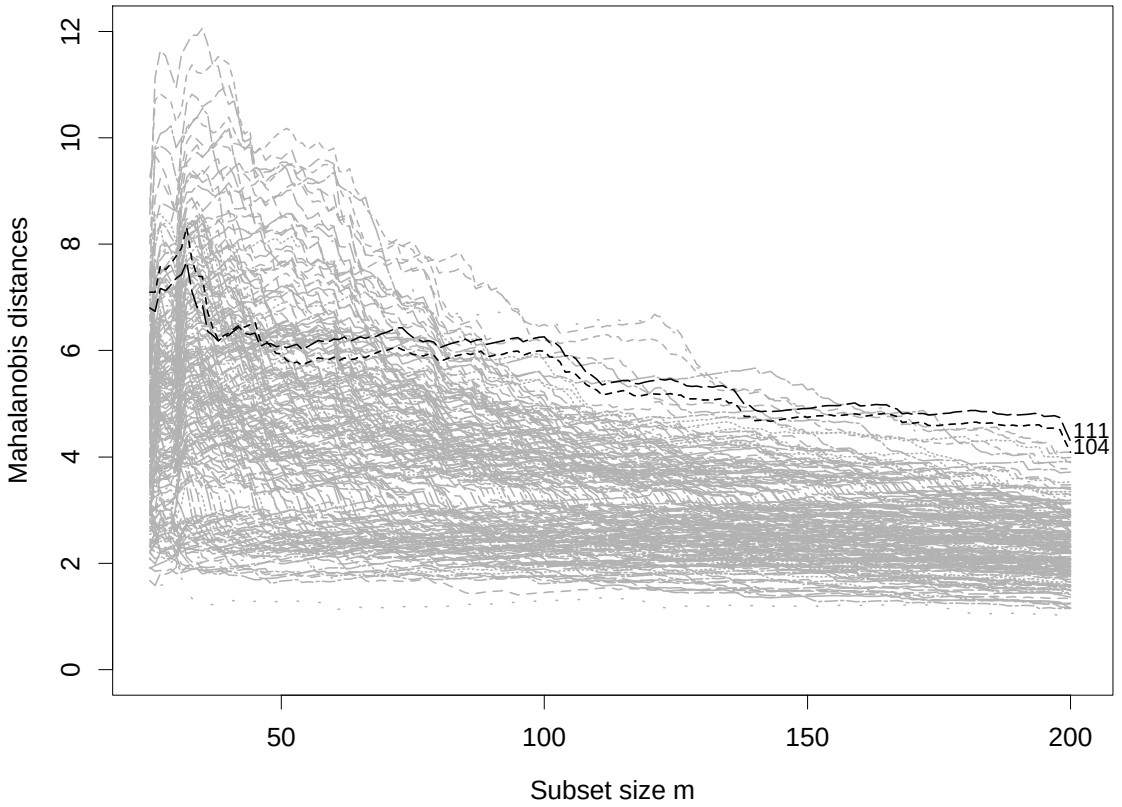
Of course, we do not have to look at a plot of all the distances



Swiss heads: forward plot of scaled Mahalanobis distances. The trajectories for units 104 and 111 are highlighted; they are initially not particularly extreme

Swiss Heads 3

The plot of unscaled distances looks similar



Swiss heads: forward plot of unscaled Mahalanobis distances. The trajectories for units 104 and 111 are again highlighted; the behaviour at the end of the search is obscured

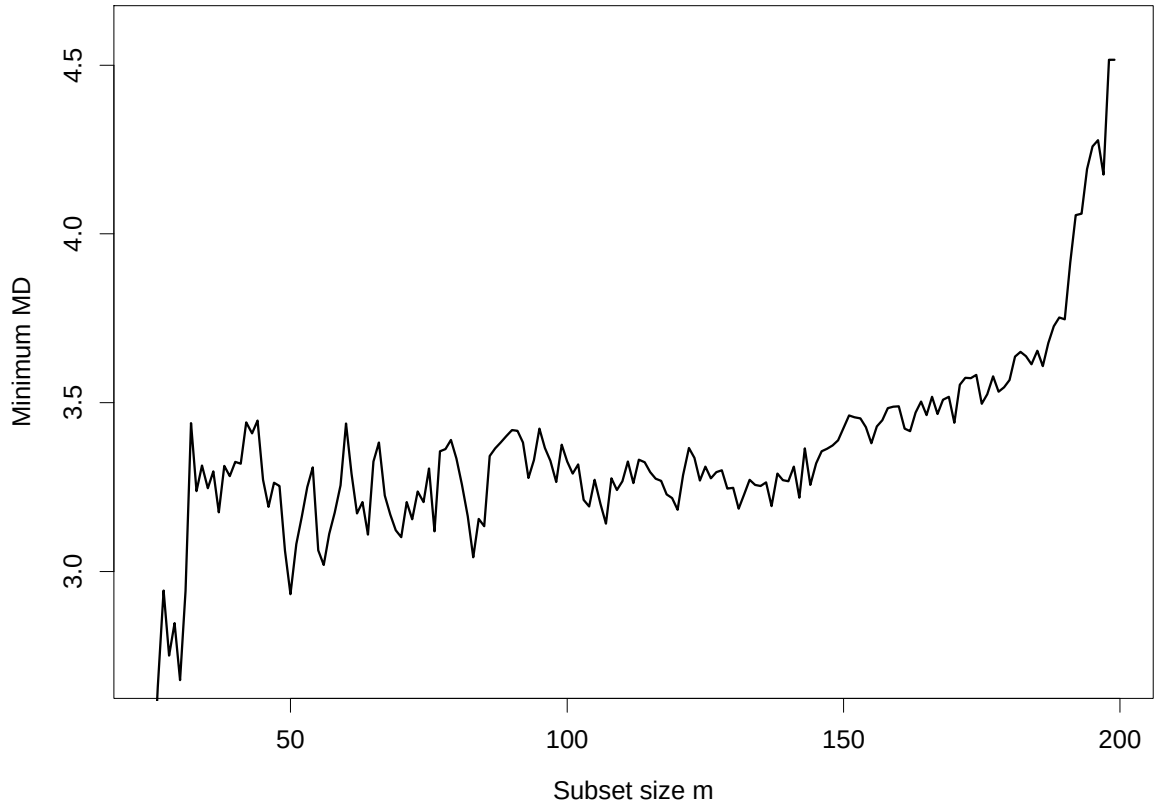
The Forward Search 4: Outliers

To detect outliers we examine the minimum Mahalanobis distance amongst observations not in the subset

$$d_{[m+1]}(m) = \min d_i(m) \quad i \notin S(m), \tag{1}$$

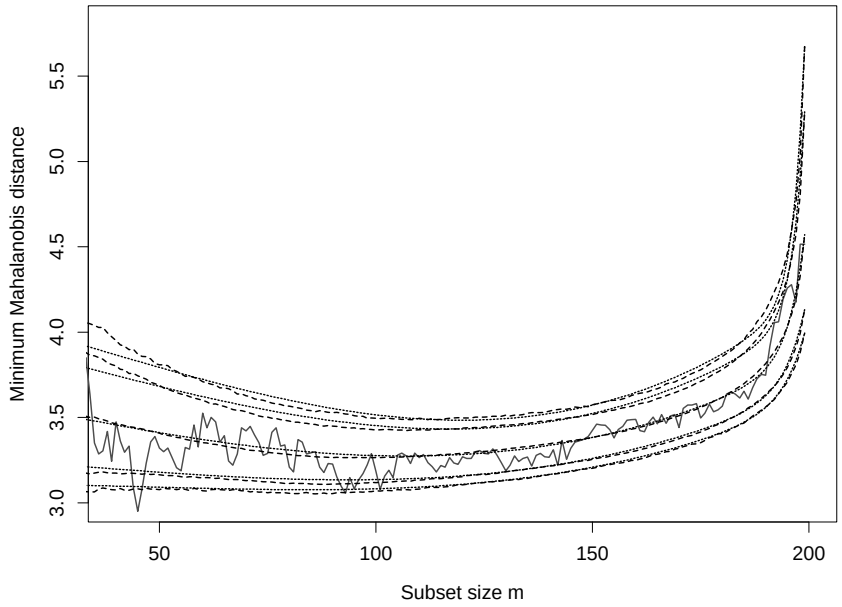
or its scaled version $d^{\mathsf{sc}}_{[m+1]}(m)$. If observation $[m+1]$ is an outlier relative to the other m observations, this distance will be large compared to the maximum Mahalanobis distance of observations in the subset.

- If observation $[m+1]$ is an outlier, so will be all the remaining $n-m-2$ observations with larger values of d_i.



- Swiss heads: forward plot of minimum distances of units not in the subset. There may be a few outliers entering at the end of the search
- Use simulation to provide distribution

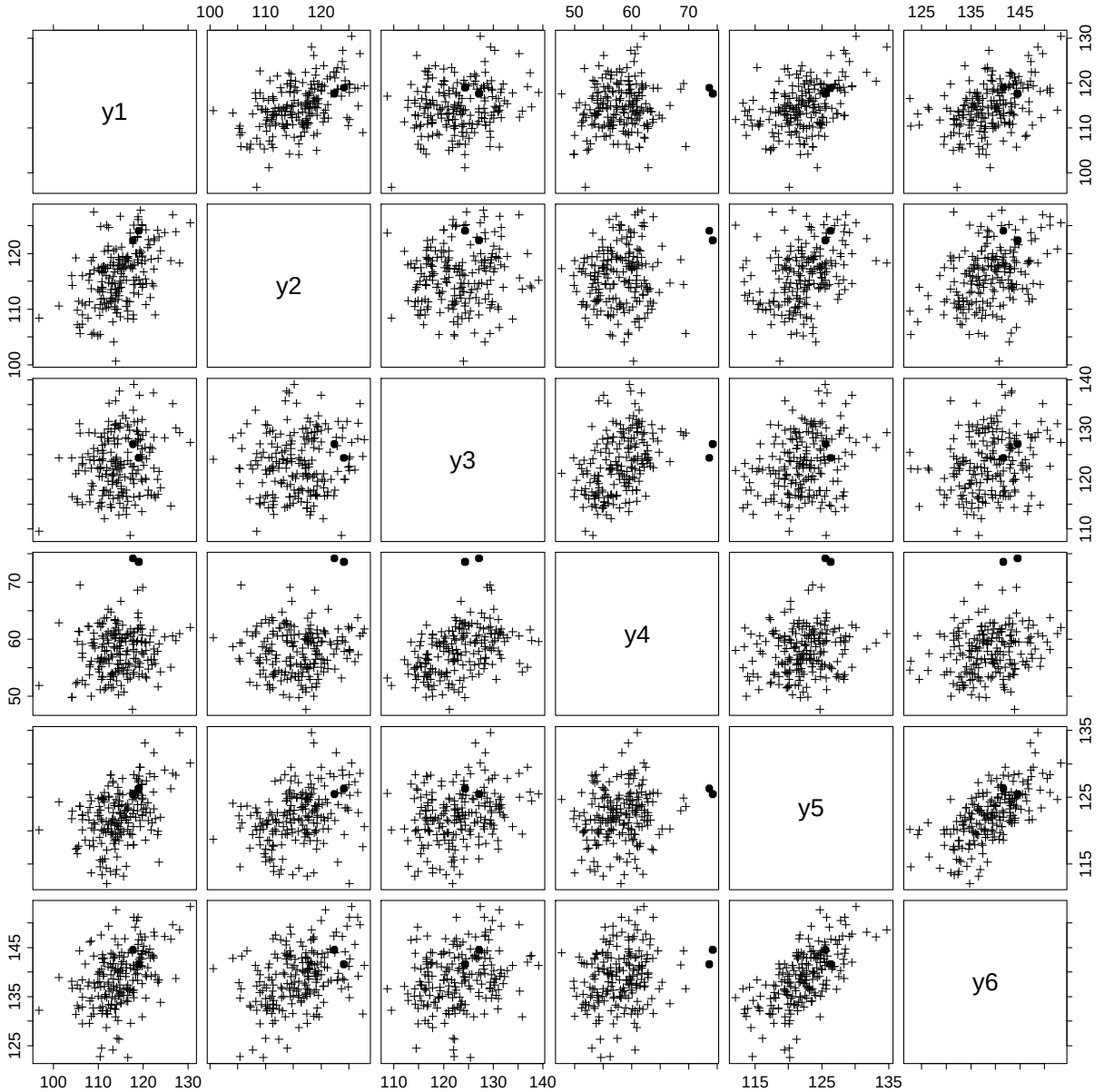
Simulation Envelopes



- Swiss heads: forward plot of minimum distances of units not in the subset.
- 1, 5, 50, 95 and 99% points of 10,000 simulation envelopes (and an approximation)
- No outliers indicated

The Forward Search 5

- Here we seem to have one normal population with two slightly extreme observations
- Do these observations matter?
  - Do they affect inferences?
  - Are they important for themselves?
- The Forward Search reduces multivariate (v-dimensional) problems to 2 dimensions
- But it may be informative to look at plots of the data in the light of the search results.



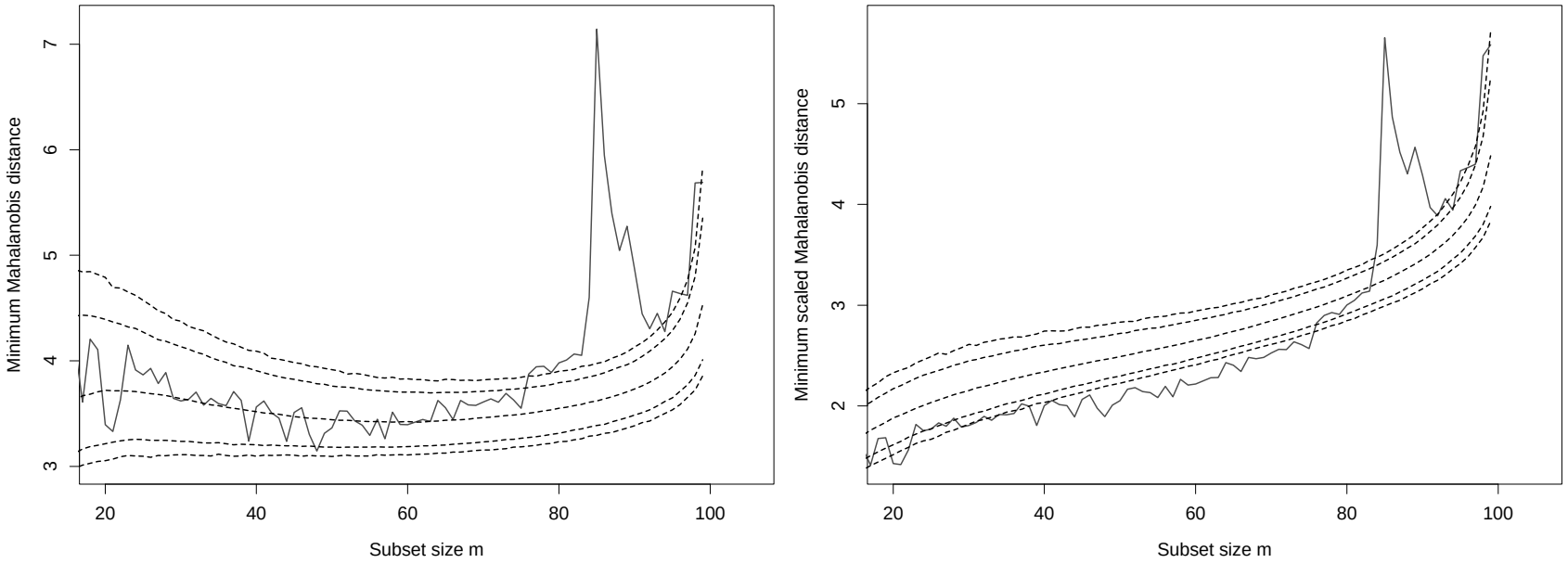
Units 104 and 111 are plotted as dots

Swiss Banknote Data

- There are (again) 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
- For the moment we just look at the forgeries

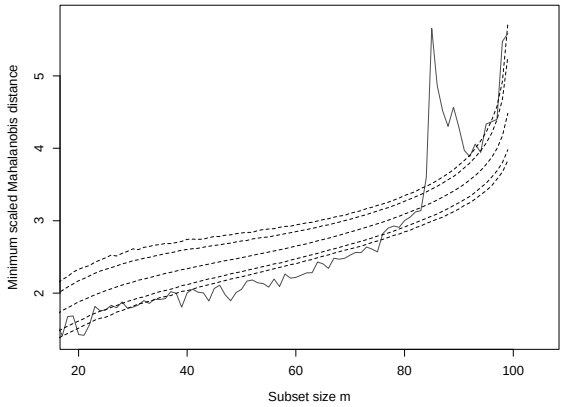
Swiss Banknote Data

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

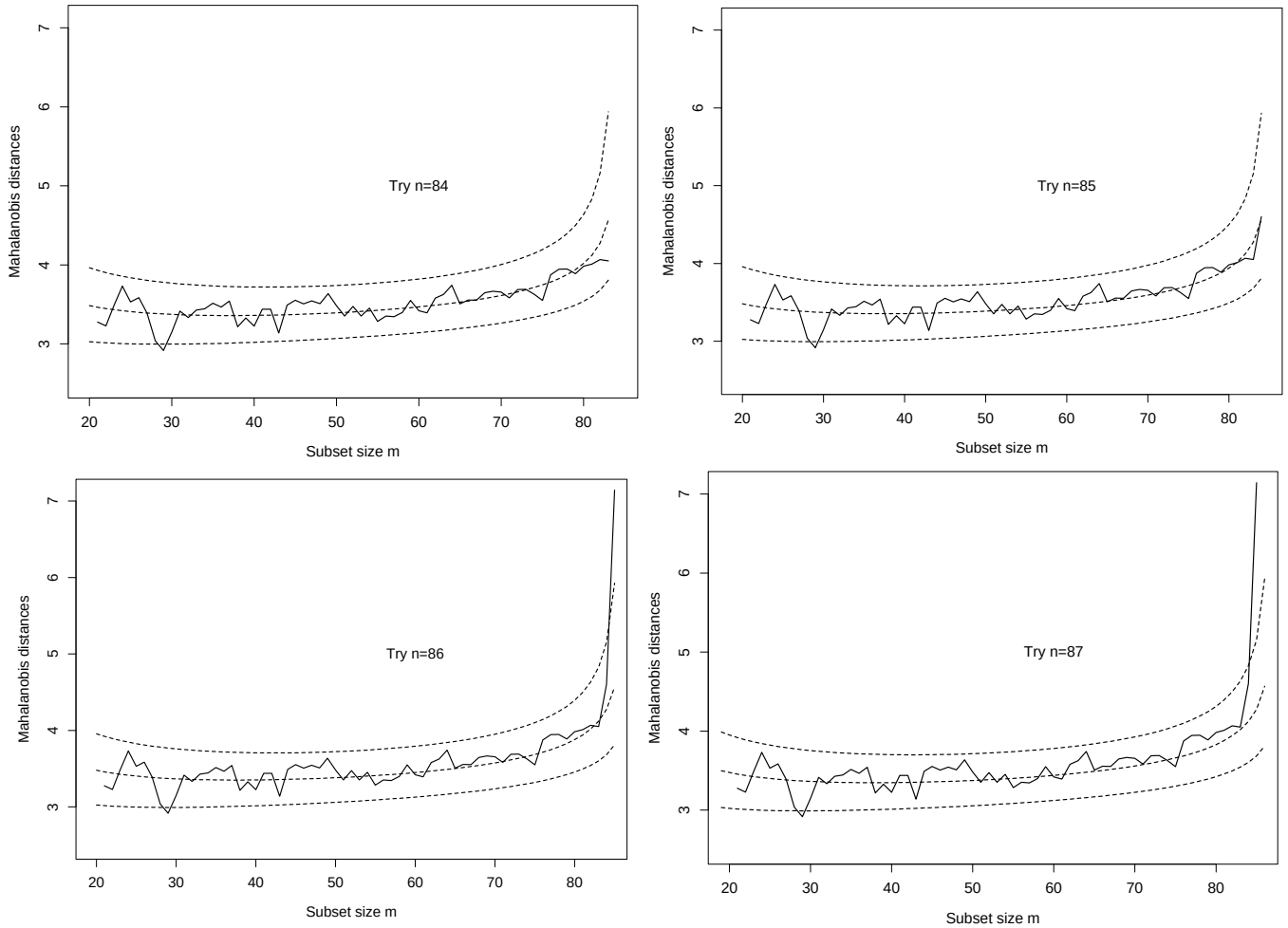


Swiss banknotes, forgeries (n = 100): forward plot of minimum Mahalanobis distance with superimposed 1, 5, 95, and 99% bootstrap envelopes using 10000 simulations. Left panel unscaled distances, right panel scaled distances. There is a clear indication of the presence of outliers which starts around $m = 84$. Note the masking

Resuperimposition



- The envelopes rise sharply at the end
- If there is masking, the final part of the search may lie inside the envelopes
- The envelopes depend on n
- We find the largest value of m for which the observed values lie within the envelopes for $n = m$



Swiss Banknotes: forward plot of minimum Mahalanobis distance. When $n = 84$ and 85, the observed curve lies within the 99% envelope, but there is clear evidence of an outlier when $n = 86$. The evidence becomes even stronger when another observation is included.

Several Populations

- In the “forgeries” there were 85 “good” observations and 15 outliers.
- Do these outliers form a group?
- What happens if there are several groups?
- For the banknote data we will at least have “genuine notes” “forgeries” and “outliers (from the forgeries)”
- Can we detect these with the FS?
- An example of a **clustering** problem with the number of groups and their properties both unknown,

References

Box, G. E. P. and D. R. Cox (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B 26*, 211–246.

Flury, B. and H. Riedwyl (1988). *Multivariate Statistics: A Practical Approach*. London: Chapman and Hall.

References

Box, G. E. P. and D. R. Cox (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B 26*, 211–246.

Flury, B. and H. Riedwyl (1988). *Multivariate Statistics: A Practical Approach*. London: Chapman and Hall.