Supplementary Material for the paper:

Wild adaptive trimming for robust estimation and cluster analysis

Andrea Cerioli
Department of Economics and Management, University of Parma, Italy
Alessio Farcomeni
Department of Public Health and Infectious Diseases
Sapienza – University of Rome, Rome, Italy
and
Marco Riani
Department of Economics and Management, University of Parma, Italy

1 Application of robust divisive clustering to the Swiss banknote data set

Flury and Riedwyl (1988, pp. 4–8) introduced 200 six-dimensional observations on Swiss banknotes withdrawn from circulation. The notes are the 1,000 Franc of the second series, the reverse of which shows a scene in a foundry (see www.snb.ch). An expert, on the basis of this scene, classified a selection of notes as genuine or forgeries. Flury and Riedwyl (1988) present measurements of aspects of the foundry scene and its relationship to the border of the note for each of 100 notes believed to be genuine and for each of 100 notes believed to be forgeries. In this data set there can be more than two groups because the fake banknotes may come from different forgers. The left panel of Figure 1 shows the monitoring of pruned trajectories of minimum Mahalanobis distances $d_{min}(m_l)$ for the whole sample of $n = 200$ banknotes and 500 random starting points. This plot clearly reveals the presence of two groups plus a set of outliers entering after step 170 when all trajectories have converged into one. The right panel, giving the scatterplot of the variables "diagonal distance" against "distance bottom border", shows the units inside the fitting subset in the step prior to the exceedance of the extreme envelope for the trajectory with the largest value of the ratio rs_t. It is apparent that in the first divisive step we can identify a group of 76 banknotes which show a distance from bottom border much greater than the others. These are clearly forged notes. Figure 2 has the same structure as Figure 1, but it refers to the subset of 124 units remaining after removal of the first tentative group. The left panel shows just one trajectory which exceeds the extreme envelope and leads to the identification of a homogenous subgroup of 93 observations. The right panel of Figure 2 shows the units forming this group, again in the scatterplot of the variables "diagonal distance" against "distance bottom border". After this second split the procedure terminates, because we do not observe any exceedance of the extreme envelope with the remaining units. We thus set $K = 2$. We conclude by noting that our tentative clustering in two groups of these data clearly separates the population of genuine notes from that of forgeries. Among the unassigned units, we find a well known groups of outliers possibly from different forgers (García-Escudero and Gordaliza, 2005; Cerioli, 2010) and some borderline observations. If we recall the procedure by which the "true" assignment were defined, there might also be a few misclassified observations. A more accurate discrimination among the two main populations, the outliers from different forgers and the possibly misclassified notes could be obtained by a refining confirmatory step, involving decreasing levels of trimming from the centroids of our tentative clustering with $K = 2$.

References

Cerioli, A. (2010). Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association 105*, 147–156.

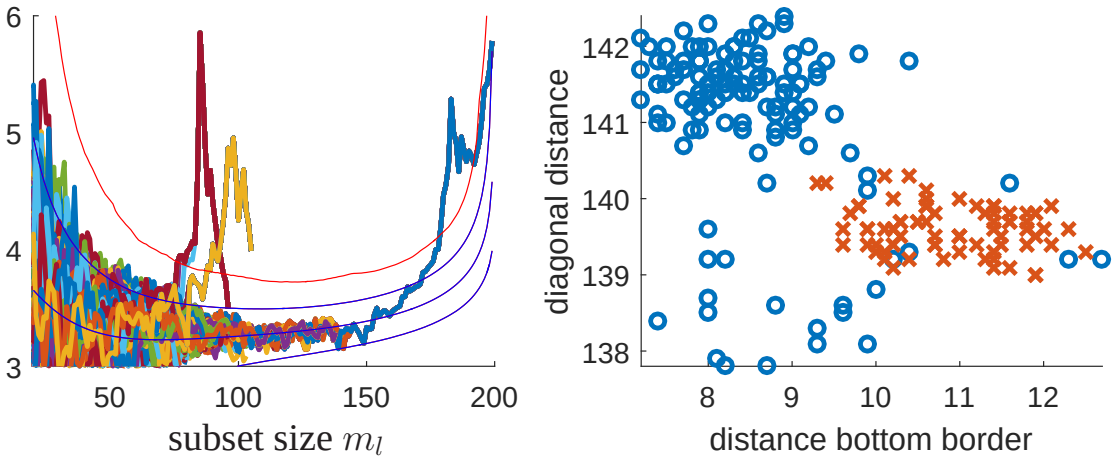

Figure 1: Swiss banknotes. Left panel: trajectories of pruned minimum Mahalanobis distances from 500 random starts with 1%, 50%, 99% and 99.9999% envelopes. Right panel: scatter plot of the variables "diagonal distance" against "distance from bottom border". Crosses correspond to the units inside the fitting subset before the first exceedance of the extreme envelope.

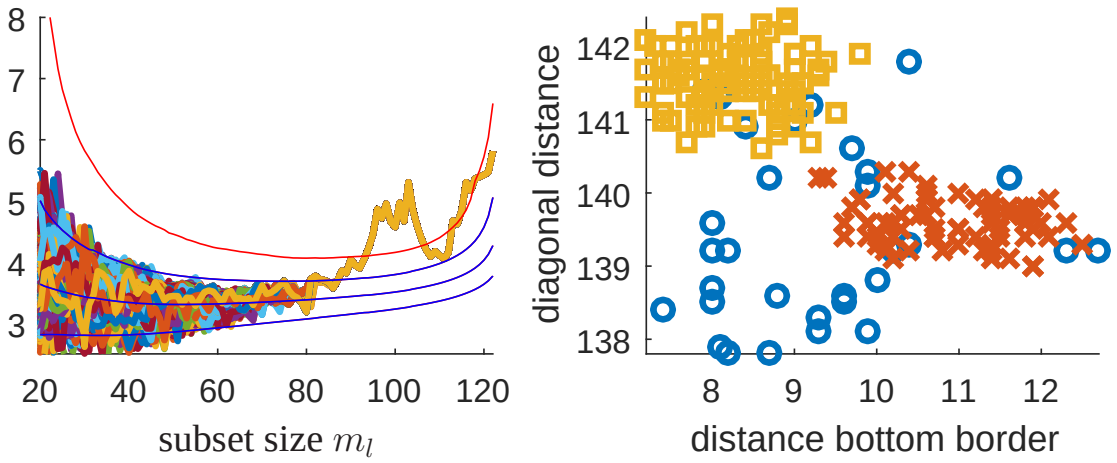

Figure 2: Swiss banknotes. As Figure 1, but now after removal of the first tentative group. In the right panel, squares correspond to the units inside the fitting subset before the first exceedance of the extreme envelope, circles to unassigned units.

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

García-Escudero, L. A. and A. Gordaliza (2005). Generalized radius processes for elliptically contoured distributions. *Journal of the American Statistical Association 100*, 1036–1045.