An Econometric Application of the Forward Search in Clustering: Robustness and Graphics

Anthony Atkinson , Marco Riani, Andrea Cerioli

Abstract: We describe a form of robust clustering using the forward search that allows the data to determine the number of clusters. We also allow for outliers and so do not force all units to be clustered. An example is given of the successful clustering of data on Italian funds.

MSC 2000: 62H30, 62-09
Key words: Mahalanobis distance, Number of clusters, Simulation envelope

1 Introduction

Classification methods are widely used in the financial services industry, for example for identifying bad credit risks and stocks with desirable (or undesirable) properties [5]. A powerful set of tools for such purposes are the statistical methods of cluster analysis. Yet, despite the sensitivity of many of these methods to outliers or incorrectly specified structures, robust statistical methods are little used. It is the purpose of the present paper to describe robust cluster analysis using the forward search. This graphics rich robust approach to clustering uses the data to identify the number of clusters, to confirm cluster membership and to detect outlying observations that do not belong to any cluster. An example is given of the analysis of some financial data on the performance of funds.

Given the poor quality of much data, the low uptake of robust methods in financial and economic areas is surprising. For example, regression is one of the major statistical tools in applied economics with ordinary least squares the preferred method of estimation. Few publications in economics apply robust methods in regression, despite the well-known susceptibility of ordinary least squares to outliers and [10] accordingly advocate the use of high-breakdown robust regression, specifically least trimmed squares, and compare robust and non-robust parameter estimates and associated t-values. These analyses are supplemented by those of [1] who, in a companion paper to this one, uses the forward search combined with graphics to reveal the structure of the data. The present paper applies similarly illuminating methods to clustering multivariate data.

Our procedure, related to robust methods for detecting outliers, uses robust Mahalanobis distances to identify and define the clusters in the data. The use of robust methods for outlier detection was popularised by [7]. A robust method with exploratory graphics for the detection of outliers in multivariate normal data is described by [9]; [8] provide a fast algorithm for the calculation of the robust distances. The robust estimates of means and covariances that we use in the calculation of

the Mahalanobis distances in this paper come from the forward search, in which subsamples of increasing size are used for parameter estimation. A wide variety of applications of the forward search in the analysis of multivariate data are given by [3]. Their analyses typically rely on inspection of forward plots of robust distances. However, in order to provide sensitive inferences about cluster membership, it is, as we show, necessary to augment such plots with envelopes of the distributions of the statistics being plotted.

Mahalanobis distances from the forward search are used by [6], in combination with envelopes, to give tests for multiple outliers in multivariate data. For the detection of outliers from one population they use single forward searches from a carefully chosen starting point. The required envelopes can be found by simulation, which will be time consuming if outliers are dropped sequentially, so that envelopes are required for a series of sample sizes, as they are in the example in our §4, where we use approximations to the envelopes derived from those in [6].

For cluster definition, as opposed to outlier identification, several searches are needed, the most informative being those that start in individual clusters and continue to add observations from the cluster until most observations in that cluster have been used in estimation. There is then a clear change in the Mahalanobis distances as units from other clusters, which are outliers from the cluster being fitted, enter the subset used for estimation. This strategy seemingly requires that we know the clusters, at least approximately, before running the searches. But we instead use many searches with random starting points to provide information on cluster existence and definition. Except at the beginning of the search, [2] show that the same set of envelopes is appropriate for searches with a random start and for those in which the starting point is selected with care.

Robust Mahalanobis distances and the forward search are defined in §2. In §3.1 we apply forward clustering to a synthetic data set with three clusters. These are recovered by our method, but missed by k-means clustering in §3.2, where use of a standard statistic indicates five groups. Our analysis of the financial data is in §4.

2 Mahalanobis Distances and the Forward Search

The main tools that we use are plots of Mahalanobis distances. The squared distances for the sample of n v-dimensional observations are defined as

$$d_i^2 = \{y_i - \hat{\mu}\}^T \hat{\Sigma}^{-1} \{y_i - \hat{\mu}\}, \tag{1}$$

where $\hat{\mu}$ and $\hat{\Sigma}$ are the unbiased moment estimators of the mean and covariance matrix of the n observations and y_i is $v \times 1$.

In the forward search the parameters μ and Σ are estimated from a subset of m observations, yielding estimates $\hat{\mu}(m)$ with $\hat{\mu}(m)_j = \bar{y}_j$ and $\hat{\Sigma}(m)$ with $\hat{\Sigma}(m)_{jk} = (y_j - \bar{y_j})^T (y_k - \bar{y_k})/(m-1)$. Here y_j and y_k are $m \times 1$. From this subset we obtain n squared Mahalanobis distances

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

To start the search for cluster identification we take a random sample of $m_0 = v + 1$ observations, the minimum size for which Σ can be estimated. We require this subset to be as small as possible to maximize the probability that all members of $S(m_0)$ come from the same cluster. This subset of m_0 observations grows in size during the search in such a way that non-cluster members will be excluded with high probability. When a subset $S(m)$ of m observations is used in fitting we order the 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 one observation, but sometimes two or more observations enter as one or more leave.

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

$$d_{\min}(m) = \min d_i(m) \quad i \notin S(m). \tag{3}$$

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. All other observations not in the subset will, by definition, have distances greater than $d_{\min}(m)$ and will therefore also be outliers.

When the observations come from a single multivariate normal population with some outliers, these outlying observations enter at the end of the search. However, if clusters are present and a search runs within a single cluster, the "outliers" will occur when the cluster is exhausted. Our bounds for a single multivariate population will then apply to this part of the search where information about clusters may be found.

3 The Detection of Clusters

We look at a synthetic examples with three clusters and some outliers. We show how random start forward searches combined with envelope plots of forward Mahalanobis distances lead to the detection of clusters. We then interrogate the forward plot to obtain an initial definition of the clusters. The example is used to describe our new procedures for cluster definition. We also compare our results with k-means clustering. We indicate how the information gained from these initial forward searches can be used as the basis of a definitive clustering.

3.1 Forward Clustering

The two-dimensional data were simulated to have three clusters of 150 slightly correlated observations with 30 lightly clustered outliers. The left-hand panel of Figure 1 shows the data; the clusters are not indicated. We used 300 random start forward searches to elucidate the structure of the data. The results are shown in the right-hand panel of the figure which is a forward plot of the values of the minimum Mahalanobis distances $d_{\min}(m)$. The forward searches fall into four classes: those

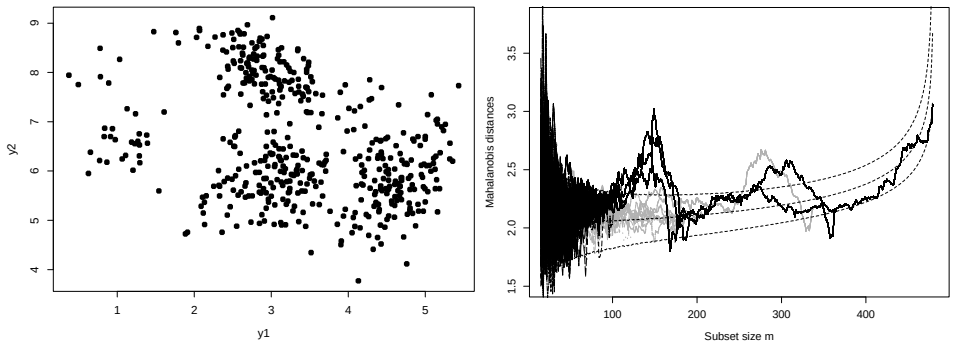

Figure 1: Example 1: three clusters and outliers. Scatter plot of the data and forward plot of minimum Mahalanobis distances from 300 random starts with 1%, 50% and 99% envelopes. Three clusters are evident around $m = 150$. Trajectories in grey always include units from both groups

that start, or soon move into, one of the three groups and those that, from the beginning of the search, include observations from more than one group. These are shown in grey on the plot. From around $m = 110$ the searches with observations from only one group start to lie systematically above the envelopes.

The shape of the envelopes is important. As the curves in Figure 1 of [2] show, the plots for smaller n and given m rise above those with larger n as m increases, the values at the end of the search hardly depending on n. The curves in Figure 1 here are behaving as those from samples of around 150. However, by $m = 180$ the three curves have returned within the envelopes as relatively remote observations from other groups enter the subsets. Shortly thereafter there are three forward plots, in which a common mean and common covariance matrix are calculated from the observations in two of the three groups. Around $m = 300$ observations from the group not so far included start to enter the subset and, by $m = 370$ a common trajectory has been established; all 300 searches have converged.

The three peaks in the right-hand panel of Figure 1 suggest that there are three clusters. We now move to cluster membership. Figure 1 shows that, in the last third of the search, all 100 searches have converged in the sense that, for each m, there is one common set of observations $S(m)$ and one value of $d_{\min}(m)$. Once two searches have the same value of $S(m)$ they will continue to follow the same path, wherever they started from, producing identical values of $d_{\min}(m)$. We can say that the *residual trajectories* of the two searches are identical. Figure 1 shows that initially there are many different values of $d_{\min}(m)$ and so many distinct residual trajectories. We now see how the number of values of $d_{\min}(m)$, and so the number of residual trajectories, declines with m.

The left-hand panel of Figure 2 shows how the number of different values of $d_{\min}(m)$ decreases with m. Initially there are 300 trajectories. The left-hand panel of the figure shows that the number decreases rapidly, reaching 1 at $m = 366$ for this

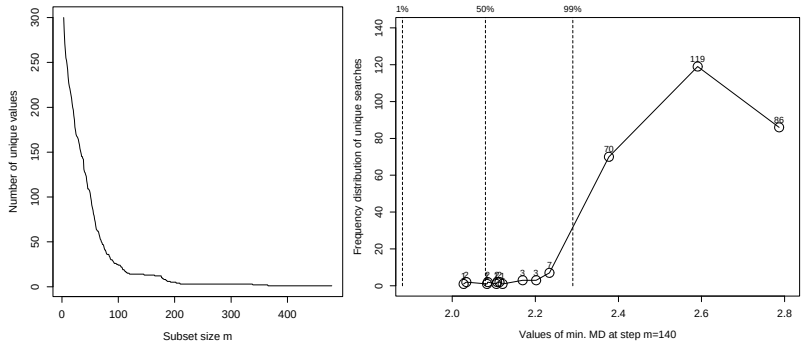

Figure 2: Example 1: number of values of $d_{\min}(m)$ and distribution of values of $d_{\min}(140)$

set of simulations. We are interested in the subsets $S(m)$ for those trajectories where there is evidence of a cluster structure. From Figure 1 this is around $m = 140$, just before the peak values, where the clusters may be beginning to be contaminated by members of other groups. To find the clusters we interrogate the right-hand panel of Figure 1 at this point to find the subsets giving rise to the larger values of $d_{\min}(140)$.

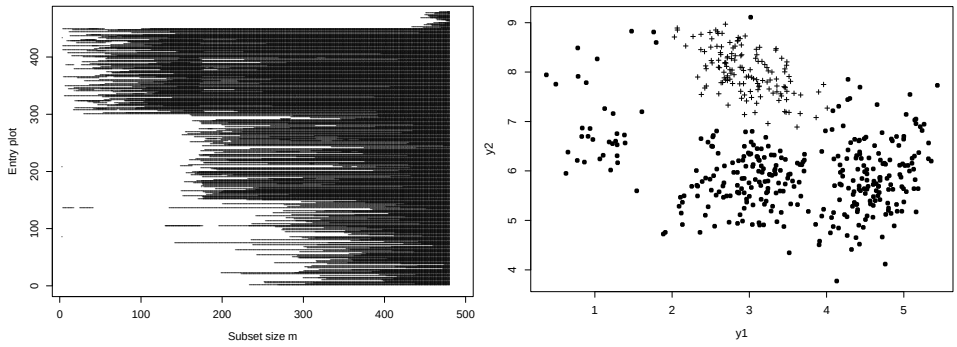

Figure 3: Example 1: entry plot and scatter plot showing identified members of the third cluster at $m = 140$

The right-hand panel of Figure 2 is a plot of the frequency distribution of the values of $d_{\min}(140)$. The vertical lines in the plot correspond to the 1%, 50% and 99% points of the envelope at $m = 140$. There are thirteen values of $d_{\min}(140)$ and so thirteen residual trajectories. The largest value occurs 86 times and the two next highest 119 and 70 times. All except 25 trajectories therefore seem to be within a cluster. The membership of the subsets for each value can be illustrated using an 'entry' plot.

The entry plot is a way of representing the membership of $S(m)$ as a search

progresses. For each m those observations included in the subset are plotted with a symbol, so that the plot becomes darker as m increases. Such plots are discussed in [3, §7.3.3]. Here we need to combine information from several searches.

The left-hand panel of Figure 3 is the entry plot for one of the 86 searches with the most extreme value of $d_{\min}(140)$ in Figure 2. Since the searches have converged at $m = 140$, all will have the same residual trajectory, so it does not matter which of the 85 we choose to plot. For $m < 140$ we select randomly from one of these searches to obtain a typical plot. Figure 3 shows clearly that we have found the trajectories that include observations from those numbered 301 - 450. The members of this cluster are unambiguously identified. The entry plot also shows that observations from the second group are next to enter the subset and that the outliers, at the top of the plot, are the last to enter. The right-hand panel of Figure 3 shows the scatter plot of the data, with the observations in the subset at $m = 140$ plotted with crosses. The search has clearly identified a meaningful cluster.

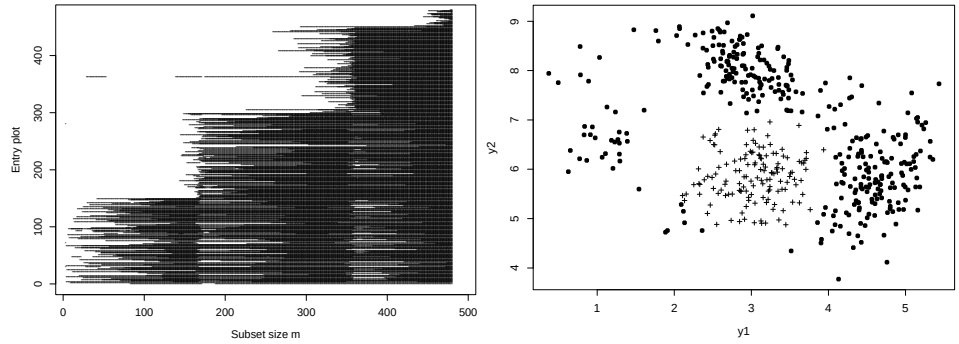

Figure 4: Example 1: entry plot and scatter plot showing identified members of the first cluster at $m = 140$

Likewise, Figure 4 shows the cluster of observations numbered 1 - 150 and their positions in the data. The third highest value of $d_{\min}(140)$ in Figure 2 gives rise to a similar entry plot, not shown here, in which observations from those numbered 151-300 have been identified as a cluster. Finally Figure 5 is for the fourth highest value of $d_{\min}(140)$ in Figure 2. Here there is no cluster structure, observations from clusters 1 and 3 being in the subsets throughout. This is in line with the interpretation of Figure 1 where the fourth highest trajectory at $m = 140$ is within the envelopes. The scatter plot matrix in the right-hand panel of Figure 5 shows that the search is identifying a long thin cluster with units from both groups and those in between. The other, lower values of $d_{\min}(140)$ in Figure 2 likewise give subsets containing units mostly from two groups.

The results shown in Figures 3, 4 and **??** clearly indicate how well our cluster identification procedure works. Since the clusters correspond to the natural order of the units, we are able to check that our method has achieved what is required. But permutation of the labels of the units, corresponding to not knowing the clus-

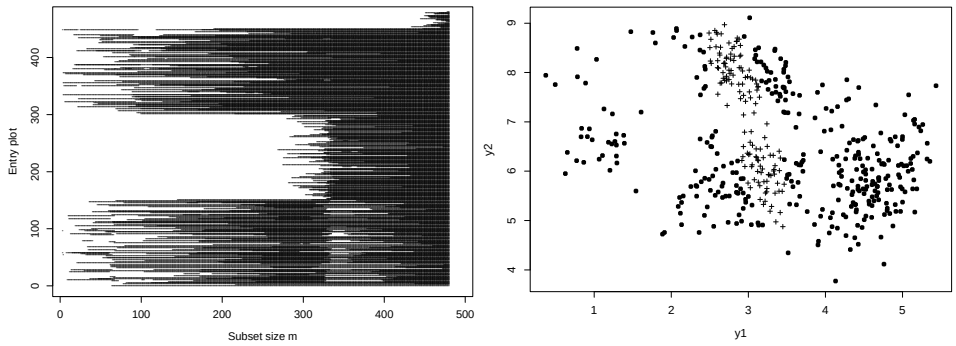

Figure 5: Example 1: entry plot and scatter plot for the fourth largest value of $d_{\min}(140)$; the search includes units from clusters 1 and 3

ter membership, would destroy this structure. Nevertheless the division of units provided by Figure 1 and the other plots provides excellent initial clusters for the further investigation demonstrated at the end of §4

3.2 Clustering with k-Means

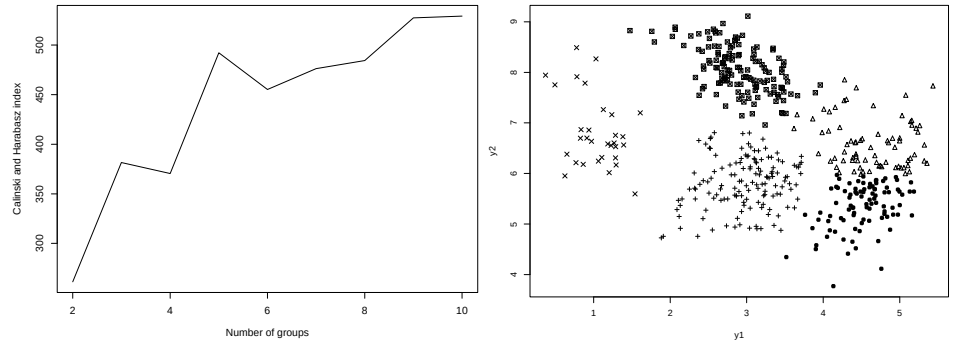

Figure 6: Example 1: clustering with k-means. Calinski-Harabasz index and the five "clusters"

Despite the clear appearance of three clusters in Figure 1, the determination of cluster numbers is a problem with many methods. As an example we apply k-means clustering to our data, without standardizing. For the selection criterion we use the Calinski-Harabasz index [4]. The left hand panel of Figure 6 shows the plot of the index against the number of groups. Large values of the index are desirable, so five groups are indicated. The right-hand panel of Figure 6 shows the scatter plot of the data with the resultant clustering. In effect, the second group, identified in Figure **??**, has been split in half and the outliers form the fifth cluster. It is hard

to know how to progress from here, whereas, with the forward search, it is natural to study the effect of the individual inclusion of each unit in a cluster.

4 Financial Data

We now apply our forward search method to the clustering of the "Financial Data" in Table A.16 of [3]. There are measurements on three variables monitoring the performance of 103 investment funds operating in Italy. The numbering of the data corresponds to two different kinds of fund. We investigate whether this information can be recovered.

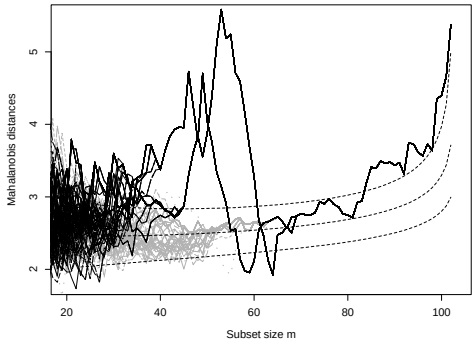

Figure 7: Financial data: forward plot of minimum Mahalanobis distances from 200 random starts with 1%, 50% and 99% envelopes. Two clusters are evident around $m = 50$

Figure 7 shows the forward Mahalanobis distances from 200 random starting points. There are two trajectories with high values around $m = 50$: one with twin peaks and one still rising. To avoid including outliers, we take our cut marginally earlier at $m = 48$. The left-hand panel of Figure 8 shows that we have two trajectories with high values and high frequencies. In the right-hand panel we give the entry plot for a search giving the highest value of $d_{\min}(48)$. The separation into two groups is virtually complete. The entry plot for the second highest value of $d_{\min}(48)$ gives a complementary division.

We now confirm this provisional clustering with single searches starting with $m_0 = v + 1$ units taken in turn from each provisional cluster. Figure 9 shows the plot for group 1 (the larger group in Figure 7) on which we have superimposed envelopes for increasing sample sizes. In the first panel $n = 52$ and all observations at the end of the search up to this value lie inside the envelopes. But for $n = 53$, in the second panel, and, even more so, $n = 54$, the last units to enter are outliers. This cluster therefore contains 52 observations. Similarly for cluster 2 we use the envelopes and find that $n = 46$. Thus, out of 103 units, 98 are clustered and 5 are remote from either cluster. Figure 10 shows the three panels of the scatter plot

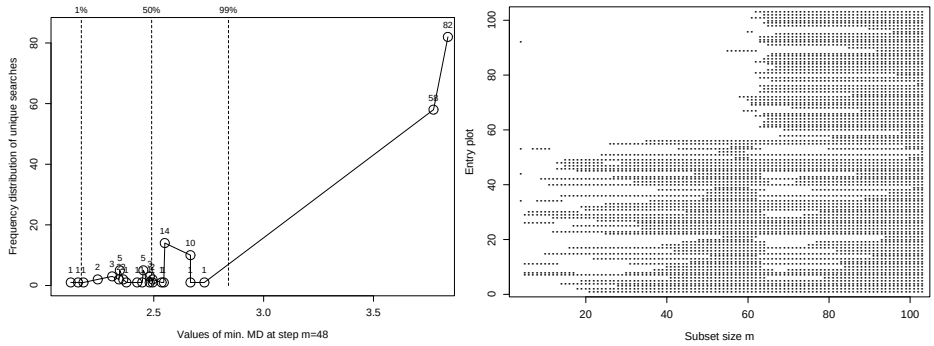

Figure 8: Financial data: distribution of values of $d_{\min}(48)$ and entry plot for the largest value of $d_{\min}(48)$. The first cluster is identified

matrix of the data, with our two clusters marked. The unclustered units are seen to lie between or remote from the two clusters.

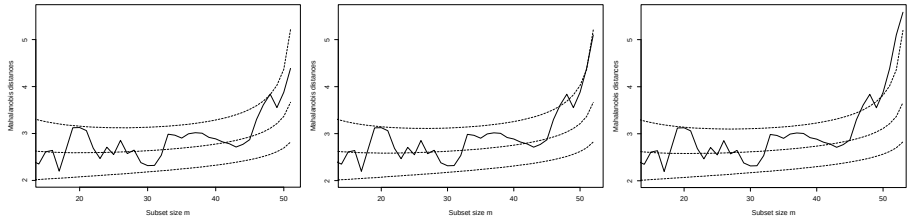

Figure 9: Financial data: forward plot of minimum Mahalanobis distances from 200 random starts with 1%, 50% and 99% envelopes for $n = 52, 53$ and 54. The cluster contains 52 observations

Our method of robust clustering has provided a simple way of clustering in which we let the data determine the number of clusters. Unlike standard methods, we are also able to allow for outliers, rather than forcing all units to be clustered. The future of our work includes automatic clustering based on extracting the information from figures such as 1 and 7.

References

[1] A. C. Atkinson. Econometric applications of the forward search in regression: robustness, diagnostics and graphics. 2006. (Submitted).

[2] A. C. Atkinson and M. Riani. Exploratory tools for defining clusters and detecting non-normality in multivariate data. 2007. (Submitted).

[3] A. C. Atkinson, M. Riani, and A. Cerioli. *Exploring Multivariate Data with the Forward Search*. Springer–Verlag, New York, 2004.

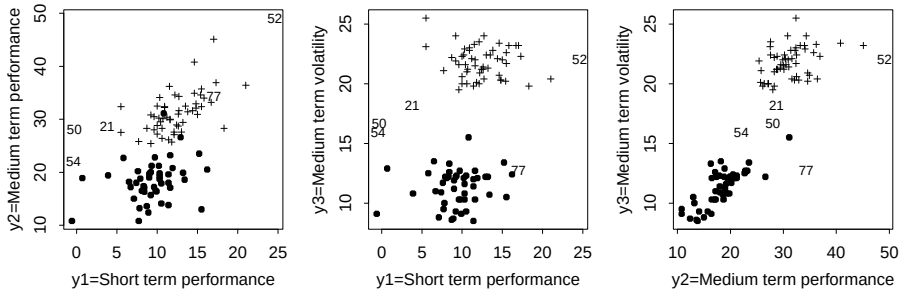

Figure 10: Financial data: the three panels of the scatter plot matrix showing cluster membership and outliers

[4] T. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3:1–27, 1974.

[5] D. J. Hand, G. Blunt, M. G. Kelly, and N. M. Adams. Data mining for fun and profit. *Statistical Science*, 15:111–131, 2000.

[6] M. Riani and A. C. Atkinson. Finding an unknown number of multivariate outliers in larger data sets. 2006. (Submitted).

[7] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. Wiley, New York, 1987.

[8] P. J. Rousseeuw and K. Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223, 1999.

[9] P. J. Rousseeuw and B. C. van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85:633–9, 1990.

[10] A. Zaman, P. J. Rousseeuw, and M. Orhan. Econometric applications of high-breakdown robust regression techniques. *Economics Letters*, 71:1–8, 2001.

Anthony Atkinson: London School of Economics, Department of Statistics, Houghton Street, London, WC2A 2AE, UK, a.c.atkinson@lse.ac.uk

Marco Riani: Università di Parma, Dipartimento di Economia, Via J. Kennedy 6, 43100 Parma, Italy, mriani@unipr.it,

Andrea Cerioli: Università di Parma, Dipartimento di Economia, Via J. Kennedy 6, 43100 Parma, Italy, statec1@ipruniv.cce.unipr.it,