Multivariate Analysis, HT2003

Visualization — Crop Viruses

© 1996 B. D. Ripley[1]

This is a dataset on 61 viruses with rod-shaped particles affecting various crops (tobacco, tomato, cucumber and others) described by Fauquet *et al.* (1988) and analysed by Eslava-Gómez (1989). There are 18 measurements on each virus, the number of amino acid residues per molecule of coat protein; the data come from a total of 26 sources. There is an existing classification by the number of RNA molecules and mode of transmission, into

39 *Tobamoviruses* with monopartite genomes spread by contact,
6 *Tobraviruses* with bipartite genomes spread by nematodes,
3 *Hordeiviruses* with tripartite genomes, transmission modeunknown and
13 'furoviruses', 12 of which are known to be spread fungally.

The question of interest to Fauquet *et al.* was whether the furoviruses form a distinct group, and they performed various multivariate analyses.

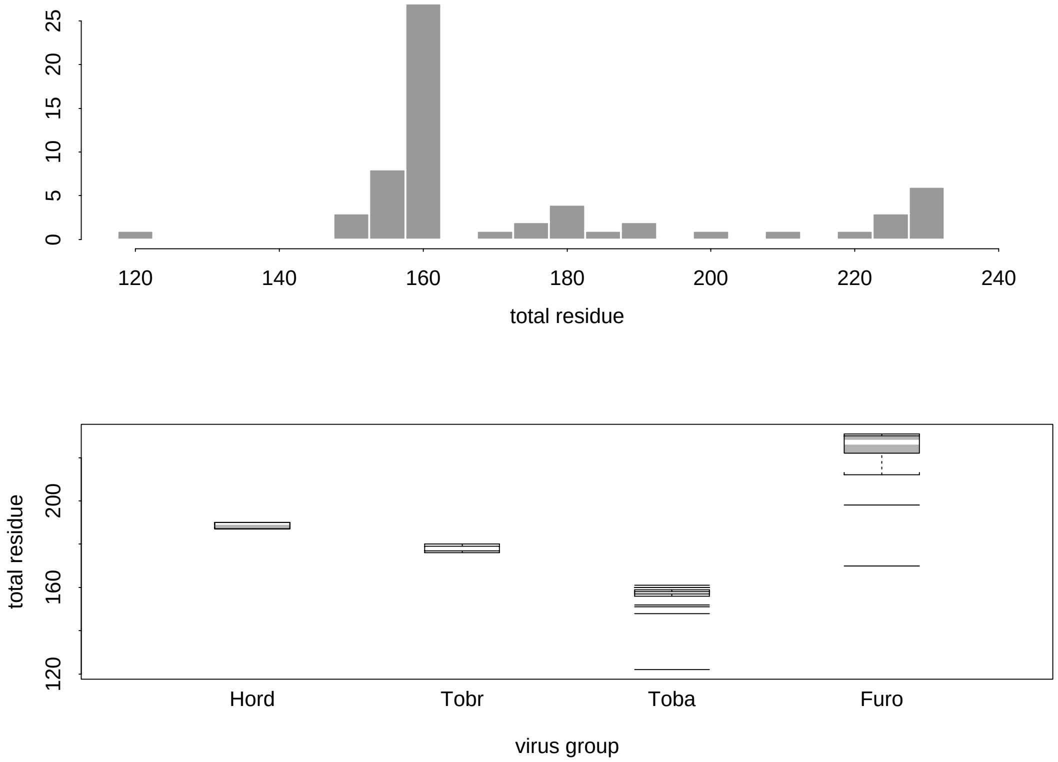


Figure 1: Histogram and boxplot by group of the viruses dataset. A boxplot is a representation of the distribution; the central grey box shows the middle 50% of the data, with median as a white bar. 'Whiskers' go out to plausible extremes, with outliers marked by bars.

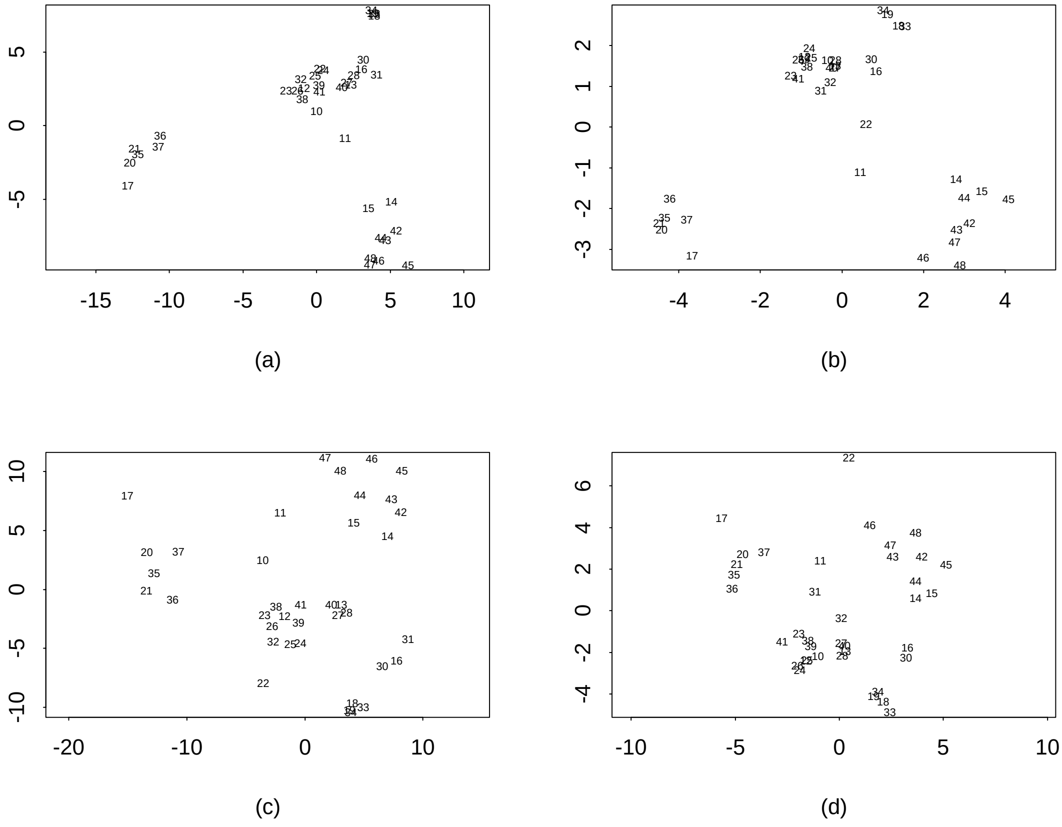
One initial question with this dataset is whether the numbers of residues are absolute or relative. The data are counts from 0 to 32, with the totals per virus varying from 122 to 231. The average numbers

[1] From *Pattern Recognition and Neural Networks* published by Cambridge University Press. The pictures will be clearer there.

for each amino acid range from 1.4 to 20.3. As a classification problem, this is very easy as Figure 1 shows. The histogram shows a multimodal distribution, and the boxplots show an almost complete separation by virus type. The only exceptional value is one virus in the furovirus group with a total of 170; this is the only virus in that group whose mode of transmission is unknown and Fauquet *et al.* (1988) suggest it has been tentatively classified as a *Tobamovirus*. The other outlier in that group (with a total of 198) is the only beet virus. The conclusions of Fauquet *et al.* may be drawn from the totals alone.

It is interesting to see if there are subgroups within the groups, so we will use this dataset to investigate further the largest group (the *Tobamoviruses*). There are two viruses with identical scores, of which only one is included in the analyses. (No analysis of these data could differentiate between the two.)

Principal Component Analysis



(a)

(b)

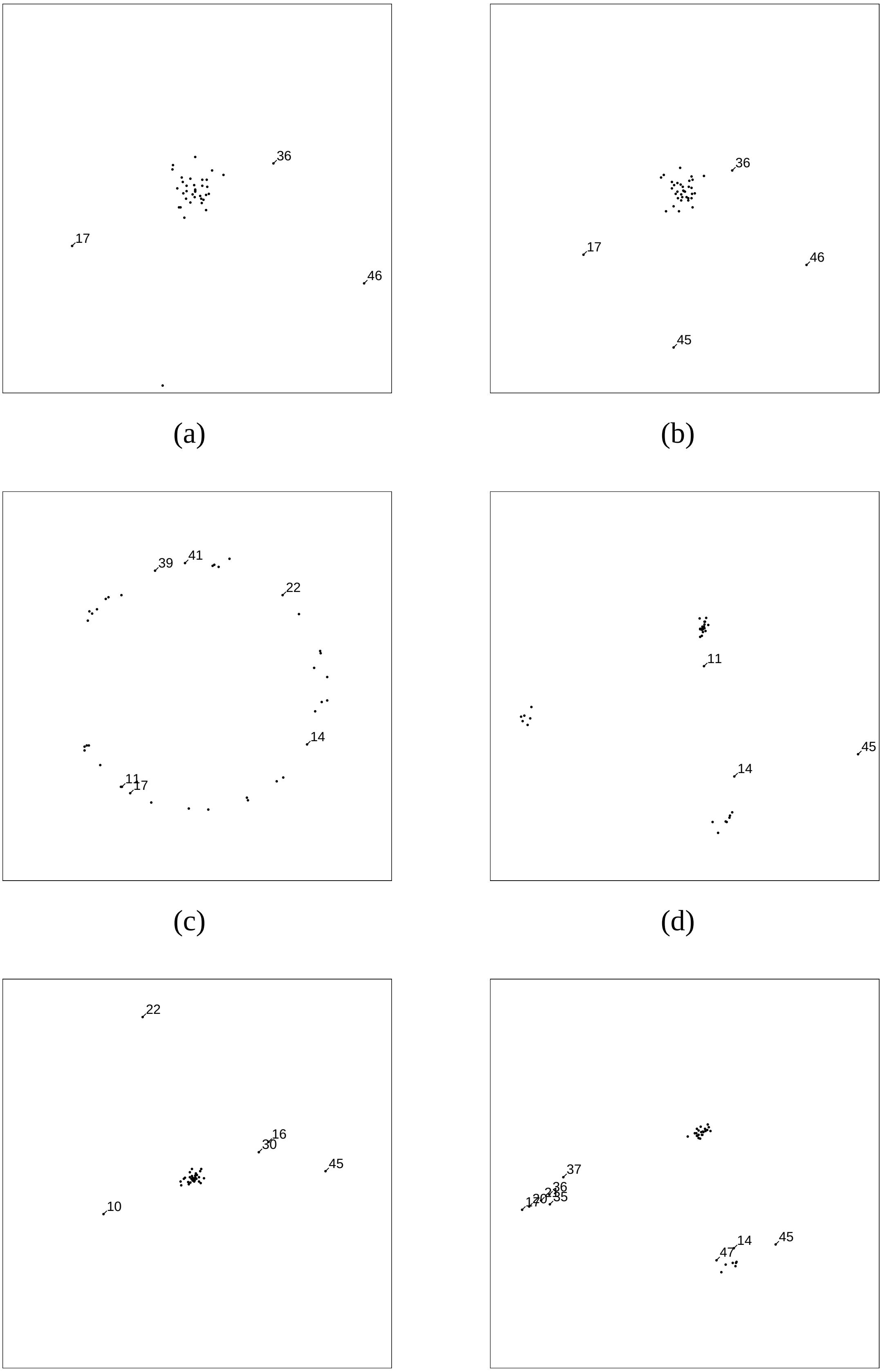
(c)

(d)

Figure 2: Principal component (top row) and Sammon mapping (bottom row) plots of the *Tobamovirus* group of the viruses example. The plots in the left column are of the raw data, those in the right column with variables rescaled to have unit variance. The points are labelled by the index number of the virus.

We consider the *Tobamovirus* group of the viruses example, which has $n = 38$ examples with $p = 18$ features. Figure 2 shows plots of the first two principal components with 'raw' and scaled variables. As the data here are counts, there are arguments for both, but principally for scaling as the counts vary in range quite considerably between variables. Virus 11 (Sammon's opuntia virus) stands out on both plots: this is the one virus with a much lower total (122 rather than 157–161). Both plots suggest three subgroupings.

In both cases the first two principal components have about equal variances, and together contain about 69% and 52% of the variance in the raw and scaled plots respectively.



(a) (b)

(c) (d)

(e) (f)

Figure 3: Projections of the *Tobamovirus* group of the viruses data found by projection pursuit. Views (a) and (b) were found using the natural Hermite index, view (c) by minimizing a_0, and views (d, e, f) were found by the Friedman–Tukey index $\int f^2$ with different choices of bandwidth for the kernel density estimator.

Projection Pursuit

Figure 3 shows six views of the main group of the viruses dataset obtained by (locally) optimizing various projection indices; this is a small subset of hundreds of views obtained in interactive experimentation in XGobi. With only 38 points in 18 dimensions, there is a lot of scope for finding a view in which an arbitrarily selected point appears as an outlier, and there is no clear sense in which this dataset contains outliers (except point 11, whose total residue is very much less than the others). When viewing rotating views of multidimensional datasets (a *grand tour* in the terminology of Asimov, 1985) true outliers are sometimes revealed by the differences in speed and direction which they display—certainly point 11 stands out in this dataset.

Not many views showed clear groupings rather than isolated points. The Friedman–Tukey index was most successful in this example. Eslava-Gómez (1989) studied all three groups (which violates the principle of removing known structure).

This example illustrates a point made by Huber (1985, §21); we need a *very* large number of points in 18 dimensions to be sure that we are not finding quirks of the sample but real features of the generating process. Thus projection pursuit may be used for hypothesis formation, but we will need independent evidence of the validity of the structure suggested by the plots.

Multidimensional Scaling

Figure 4 shows a local minimum for ordinal multidimensional scaling for the scaled viruses data. (The fit is poor, with $STRESS \approx 17\%$, and we found several local minima differing in where the outliers were placed.) This fit is similar to that by Sammon mapping in Figure 2, but the subgroups are more clearly separated, and viruses 10 (frangipani mosaic virus), 17 (cucumber green mottle mosaic virus) and 31 (pepper mild mottle virus) have been clearly separated. Figure 5 shows the distortions of the distances produced by the Sammon and ordinal scaling methods. Both show a tendency to increase large distances relative to short ones for this dataset, and both have considerable scatter.

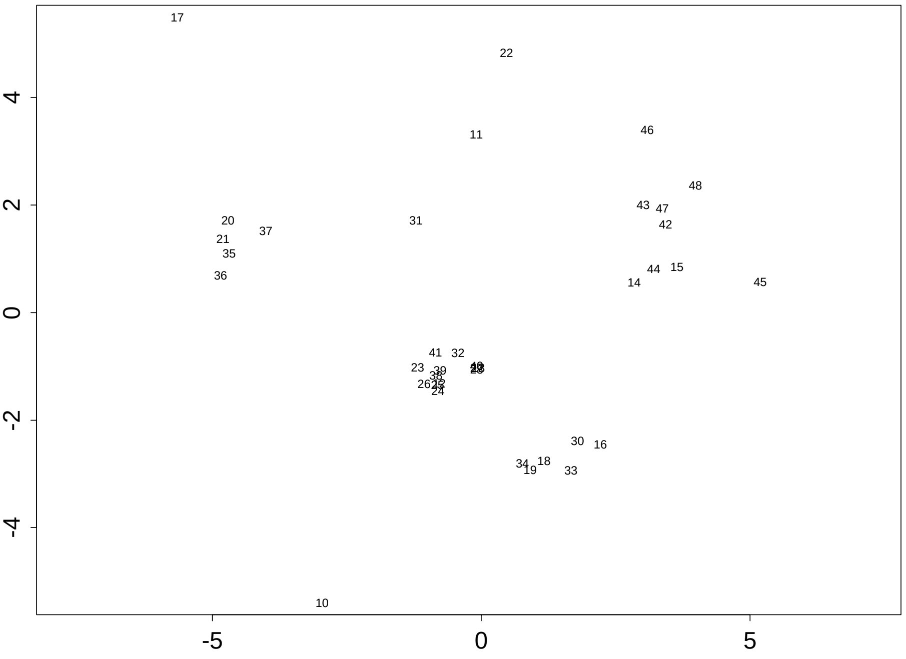


Figure 4: Non-metric multidimensional scaling plot of the *Tobamovirus* group of the viruses example. The variables were scaled before Euclidean distance was used. The points are labelled by the index number of the virus.

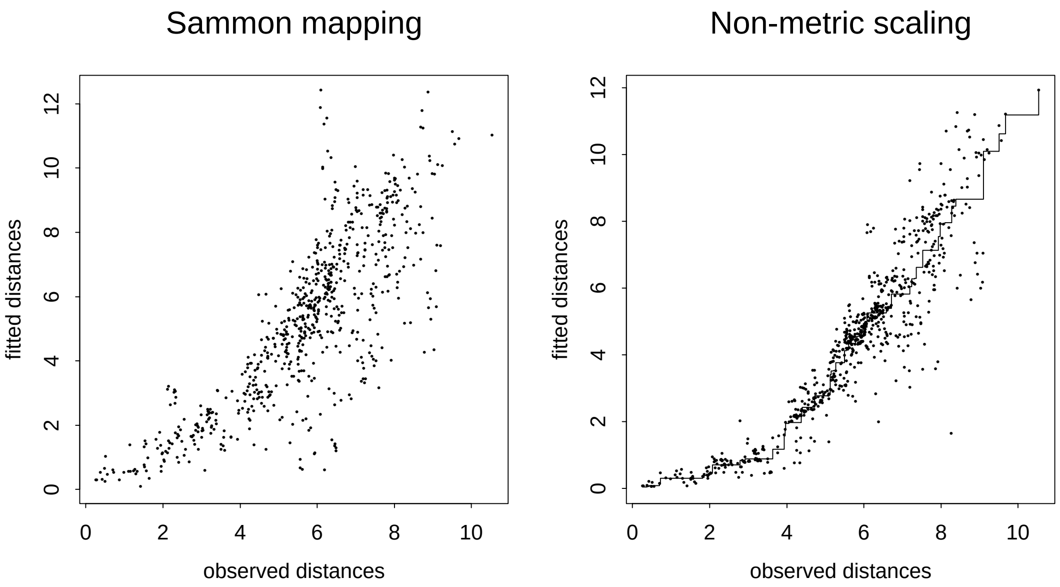


Figure 5: Distortion plots of Sammon mapping and non-metric multidimensional scaling for the viruses data. For the right-hand plot the fitted isotonic regression is shown as a step function.

Figure 4 shows some interpretable groupings. That on the upper left is the cucumber green mottle virus, the upper right group is the ribgrass mosaic virus and two others, and a group at bottom centre-right (16, 18, 19, 30, 33, 34) are the tobacco mild green mosaic and odontoglossum ringspot viruses.

Cluster Analysis

Figure 6 shows the clusters for 6-means for the virus data. The iterative process has to be started somewhere, and in this case was initialized from a hierarchical clustering discussed below. The choice of 6 clusters was by inspection of the visualization plots discussed above and the dendrograms shown in Figure 7.

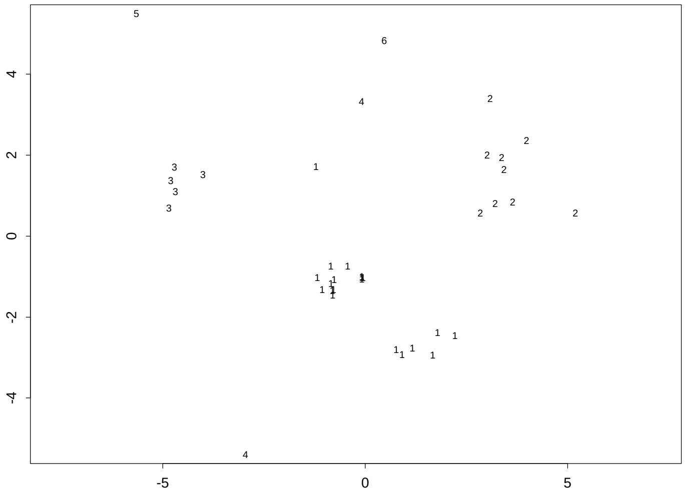
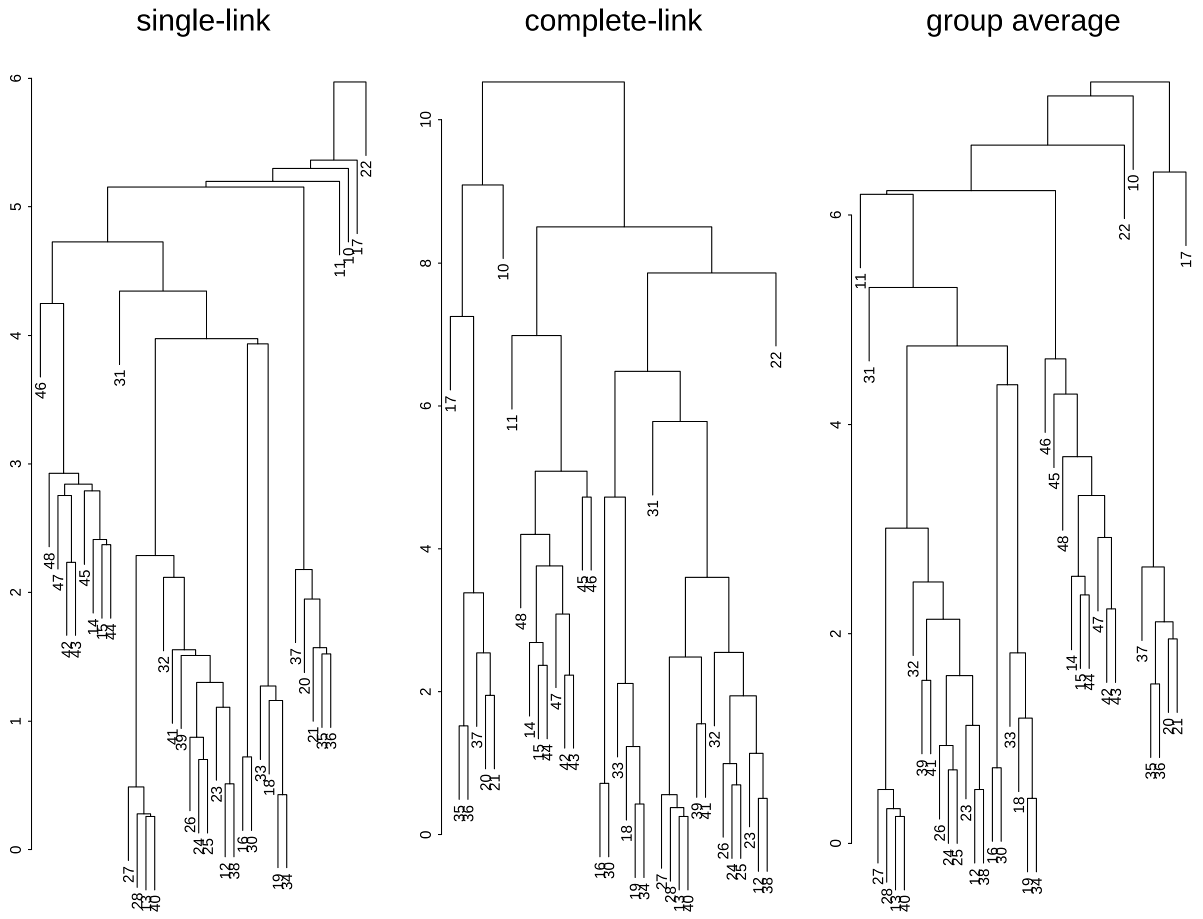


Figure 6: The clusters suggested by k-means for $k = 6$ for the virus data displayed on the ordinal multidimensional scaling plot.



single-link

complete-link

group average

Figure 7: Dendrograms from three common hierarchical clustering techniques applied to the scaled viruses data. Such plots show the dissimilarity at which clusters are merged on the vertical scale and so show the construction process from bottom to top.

Figure 7 shows dendrograms produced by single-link, complete-link and group-average clustering for the viruses data. All identify viruses 10, 11, 17, 22 and 31 as loosely connected to the rest, and single-link also highlights virus 46. (We note that 10, 11, 17, 31, 46 and 48 are called ‘miscellaneous’ in the original source.) Nevertheless, each graph gives the impression of three or four major groupings of viruses.

References

Asimov, D. (1985) The grand tour: a tool for viewing multidimensional data. *SIAM Journal on Scientific and Statistical Computing* **6**, 128–143.

Eslava-Gómez, G. (1989) *Projection Pursuit and Other Graphical Methods for Multivariate Data.* Unpublished D. Phil thesis, University of Oxford.

Fauquet, C., Desbois, D., Fargette, D. & Vidal, G. (1988) Classification of furoviruses based on the amino acid composition of their coat proteins. In *Viruses with Fungal Vectors* eds J. I. Cooper & M. J. C. Asher, pp. 19–38. Edinburgh: Association of Applied Biologists.

Huber, P. J. (1985) Projection pursuit (with discussion). *Annals of Statistics* **13**, 435–525.