Robustness aspects of model based cluster analysis

C. Hennig

[1] ETH Zürich, Seminar für Statistik, CH-8092 Zürich

Keywords: Normal mixtures, mixtures of t-distributions, noise component, number of clusters

1 Improving the robustness of Normal mixture fitting

ML-estimation based on mixtures of Normal distributions offers a flexible tool for cluster analysis. It suffers from certain robustness problems, as can be expected: A single outlier will break down at least the parameter estimators of one of the mixture components. Some ideas to overcome such problems will be presented. Here are the two best known approaches:

- The software `mclust` (Fraley and Raftery (1998)) allows the addition of a mixture component accounting for "noise", modeled as a Poisson process on the convex hull of the data.
- The software `EMMIX` (McLachlan and Peel (2000)) can be used to fit a mixture of multivariate t-distributions instead of Normals.

Both approaches were successfully applied to a couple of examples, but their robustness is not systematically established up to now.

2 Robustness measures in cluster analysis

Robustness measures in cluster analysis should characterize the most relevant robustness properties of a method, and it should be possible to evaluate them for the methods of interest. It seems that these two goals are difficult to attain at the same time. Some aspects of defining suitable a breakdown point for cluster analysis will be discussed:

- Breakdown could be defined in terms of parameters (Garcia-Escudero and Gordaliza (1999)) or in terms of the classification of the points (Kharin (1996)) with differing results.
- Breakdown properties in cluster analysis will always depend on the constellation of the "good" clusters. For example, Garcia-Escudero and Gordaliza (1999) report that even robust methods break down in situations where a fixed number of clusters is inadequately specified.
- A robust method for cluster analysis should be able to estimate the number of clusters as well (as `mclust` and `EMMIX` do). A breakdown measure for such a situation will be proposed.

3 Robustness properties of mclust and EMMIX

The robustness of model based mixture methods depends on the implementation, in particular on the initialization of the EM-algorithm and on the ability to cope with clusters with singular covariance matrix. Some experiments and result will be presented.

References

C. Fraley and A.E. Raftery (1998). How many clusters? Which clustering method? Answers via model based cluster analysis. *Computer Journal*, 41, 578–588.

L.A. Garcia-Escudero and A. Gordaliza (1999). Robustness properties of k means and trimmed k means. *Journal of the American Statistical Association*, 94, 956–969.

Y. Kharin (1996). *Robustness in Statistical Pattern Recognition.* Kluwer, Dordrecht.

G. McLachlan and D. Peel (2000). *Finite Mixture Models.* Wiley, New York.

Please fill in this form and mail it together with your abstract.

My abstract fits best to topic number 14. (**“PLEASE INDICATE HERE THE NUMBER FROM THE LIST OF TOPICS BELOW WHICH BEST FITS TO YOUR ABSTRACT”**.).

List of Topics:

1. Algorithms
2. Applications
3. Biostatistics
4. Computing and graphics
5. Data analysis
6. Data mining
7. Economics, finance
8. Efficiency and robustness
9. Functionals and bias
10. Fuzzy statistics
11. Geostatistics
12. Inference for robust methods, model testing
13. Location depth and regression depth
14. Multivariate methods
15. Neural networks
16. Rank-based methods
17. Regression quantiles, trimming
18. Robust covariance
19. Robust designs
20. Robust regression
21. Time series analysis
22. Wavelets
23. Other (please specify)