Using Finite Mixtures to Robustify Statistical Models

M.A. Jorgensen

[1] Department of Statistics, University of Waikato, Private Bag 3105, Hamilton, New Zealand

Keywords: Finite Mixture Model, EM algorithm, Influence, Robust Regression

1 Introduction

Finite mixture models such as $0.98N(\mu, \sigma^2) + 0.02N(\mu, 3^2\sigma^2)$ are common in the literature but usually as a model for the generation of contaminated data rather than for the analysis of such. We propose the latter role for finite mixture models in this talk. Specifically we propose as candidates for robust estimators the maximum likelihood estimators of the 'regular' component of a two component mixture model. The regular component would be a normal (nonrobust) statistical model and the other component would be a dispersed, fully specified, component intended to model - or at least mop up - the outlying data.

We discuss a method for calculating the influence curve for such parameter estimates and illustrate it with influence function calculations in the context of the univariate location/spread estimation and robust regression.

2 One-step influence functions

The computation of influence functions for maximum likelihood estimators is complicated by the fact that they are normally calculated by iterative algorithms. One work-around is to use a functional describing a single step of an algorithm as a surrogate for the true functional describing the estimator. The one-step influence function is the influence function of this 'single-step' functional. Jorgensen (1993) shows that this does yield the true influence function when the iterative algorithm employed is Newton's method. However mixture models are usually fit by the EM algorithm (McLachlan and Krishnan, 1997) for which this approximation is much less satisfactory. Jorgensen (1993) gives a relationship between the one-step and true influence curves which is our main tool in the present investigation.

3 Case studies

We look in particular at two simple situations:

3.1 Location/Spread

We calculate the influence functions for $\hat{\mu}$, $\hat{\sigma}^2$ and $\hat{\pi}$ in the model $(1-\pi)N(\mu, \sigma^2) + \pi N(\nu_0, \tau_0^2)$ where ν_0 and τ_0 are fixed constants.

3.2 Simple Regression

We calculate the influence functions for $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$ and$\hat{\pi}$ in mixture models in which $Y_i|X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$ in the 'regular' component and the outlier component is a dispersed circular bivariate normal. The effect of the choice for the marginal distribution of X will be discussed.

We will illustrate with some data sets.

4 Breakdown

The investigation the breakdown points of these estimators is complicated by the need to specify starting values. Nevertheless we will give heuristic arguments to suggest that the estimators perform very well when faced with dispersed contamination, but are likely to be upset by a compact group of outliers.

We conclude with speculations on the kind of situations where this approach to robust estimation is likely to be useful.

References

M.A. Jorgensen (1993). Influence functions for iteratively defined statistics. *Biometrika*, 80, 253–265.

G.J. McLachlan and T. Krishnan (1997). *The EM Algorithm and Extensions.* Wiley-Interscience, New York.

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

My abstract fits best to topic number 20. Or more exactly 23 (Robustification of general statistical models.)

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)