Dankmar Böhning
Institute for Social Medicine, Epidemiology, and Health Economy
Charité FU/HU Berlin, Germany

Discussion of the Presentations by

1. Murray Aitkin
2. Christine McLaren
3. Jiahua Chen

16. July 2004

Session on
Mixture Modelling in Biometrics
Organizer: Geoff McLachlan
IBC, Cairns, Australia, 2004

Some General Remarks

- all are based upon (sometimes several) publications in prestiguous journals
- cover different aspects:
  - mixture modeling for longitudinal, binary data (Aitkin)
  - normal mixture models applied to identify transferrin saturation in large screening studies (McLaren)
  - distributional theory for the likelihood ratio statistic in mixture models (Chen) ($k = 1$ vs. $k > 1$ and $k = 2$ vs. $k > 2$)
- motivated by different interests
- all are very interesting

1. Murray Aitkin

presentation discusses modeling of longitudinal, binary data (with covariates) using

- multivariate normal distribution for random coefficient vector (integral is approximated with Gaussian Quadrature)
- non-parametric distribution for random coefficient vector (NPMLE estimated via the EM algorithm)
- which are generalized to include a more explicit modelling of the serial dependence using a transitional GLM under incorpoaration of unobserved heterogeneity (by means of GQ or NPMLE)

Questions

- How can the GQ/NPMLE-model and the autoregressive model be discriminated?
- In particular, looking at the child obesity data, the first analysis provides support for the autoregressive model (child obesity strongly influences later development) which largely disappears when including covariates, in particular, gender.
- Mixture likelihoods are often flat - with specific consequences (EM slow, stopped too early, wide CIs,...). I expect this to be the case here, in particular. Experiences?

2. Christine McLaren

presentation discusses modeling of transferrin saturation (TS) with mixtures of normals (mixing is on both, mean and variance parameter) in two large screening studies (HEIRS, 10,000 men/15,000 women), (Kaiser Permanente Data, 14,000 men/ 14,000 women).

- results provide evidence in both studies for a 3 component mixture of normals
- first component receives large weight, third component tiny, means clearly separated, first two component variances close
- similar results in two independent studies supports a more general finding here

Questions

- Methods for testing various models M_k :

  $$p_1 N(\mu_1, \sigma_1^2) + p_2 N(\mu_2, \sigma_2^2) + ... + p_k N(\mu_k, \sigma_k^2)$$

  have been provided (based on resampling the LRS under M_k while testing vs. M_{k+1}) and concluded that $k = 3$ is giving the best fit based on comparing M_1 vs. M_2 and M_2 vs. M_3.

  - What about M_3 vs. M_4?
  - And, more importantly, does M_3 provide a good fit to the data (as measured for example with the KS-statisitic)?

- $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are very similar in all cases. Have you looked at

  $p_1 N(\mu_1, \sigma^2) + p_2 N(\mu_2, \sigma^2) + p_3 N(\mu_3, \sigma_3^2)$ **vs.**

  $p_1 N(\mu_1, \sigma_1^2) + p_2 N(\mu_2, \sigma_2^2) + p_3 N(\mu_3, \sigma_3^2)$

- finally, some other model, the log-normal, say, might provide a similar GOF. How to discriminate between the two?

3. Jiahua Chen

presentation discusses likelihood ratio testing for mixtures with some general component density $f(x, \theta)$ leading to a mixture

$$\sum_{j=1}^{k} p_j f(x, \theta_j)$$

and provides results for testing

$$\text{Part I: } k = 1 \text{ vs. } k > 1$$

and

$$\text{Part II: } k = 2 \text{ vs. } k > 2$$

- reviews some previous results on LRT distribution (difficult to use)
- and suggests to use the *penalized* log-likelihood

$$pl_n(G) = \sum_i \log f(x_i, G) + C \sum_{j=1}^{k} \log p_j$$

- MPLEs $\hat{G}_1$ and $\hat{G}_0$ of G_1 and G_0 are then used in the (conventional) LRT

- Result (under H_0 and regularity conditions)

$$LRT \to 0.5\chi_0^2 + 0.5\chi_1^2$$

Part I: Questions

- The asymptotic result is nice and connects to previous results. Is the result independent of the penalty parameter $C > 0$?
- How much complication arise due to using pl_n instead of l_n? Is there a (modified) EM, and if, how many complications?

coming to Part II:

- for testing $k = 2$ vs. $k > 2$ more complications arise:
- when finding $\hat{G}_1$ the restriction $k \geq k^* = max\{[1.5/\pi_0], [1.5/(1-\pi_0)], 4\} \geq 4$ occurs.
- Result (under H_0, $k \geq k^*$ and regularity conditions):

$$LRT \rightarrow \left(0.5 - \frac{\alpha}{2\pi}\right)\chi_0^2 + 0.5\chi_1^2 + \frac{\alpha}{2\pi}\chi_2^2$$

Part II: Questions

- A question of understanding: is the condition $k \geq k^*$ only for estimating G_1 (this is what I think) or is it a restriction on the alternative?

- The asymptotic result is nice, though less nice than the previous one, since it involves the parameter α which seems difficult to determine. Is there another use of the result than as a benchmark for simulation studies?
- And is not here the same critique appropriate that was mentioned in the presentation with regard to the previously existing results (namely, that they are difficult to use)?