amc technical brief

Editor: Michael Thompson Analytical Methods Committee AMCTB No 23 March 2006

Mixture models for describing multimodal data

An essential precept of statistics is that we always look at a graphical presentation of our data before calculating summary statistics or drawing any inference. Sometimes the inference is blazingly obvious without statistics. Sometimes, however, we see immediately that the data are inconsistent with the model, the assumptions such as the normal distribution that underlie the inference we wish to draw. In such cases we have to make alternative assumptions and use different, often more complicated, statistical methods. A common occurrence of the type arises when the dataset is (or appears to be) multimodal (Figures 1, 2). To handle that, we can often use mixture models to provide a useful summary of the data.

Multimodal datasets can arise when results from two or more different processes are combined. For example, results obtained by participants in a proficiency test when two or more inconsistent analytical methods are in use (Figure 1), or when objects from two different sources are analysed (Figure 2). When there is additional information defining such grouping, or the results are completely resolved into groups, we can treat the data subsets separately. When there is no such information or resolution, it's often worth considering the use of mixture models.

Mixture models simply assume that the dataset is a mixture, in unknown proportions, of results from two or more different unknown populations. We don't *have* to assume that the populations are normal distributions, but that assumption is usually plausible and useful. The computing task is then finding estimates of the means and standard deviations of the component distributions and the proportion in which they are present. A very efficient way to do that is to use the 'EM algorithm', which employs maximum likelihood estimation. An example is shown in Figure 3. This method is easy to encode and quick to run (see the brief introduction overleaf). But you don't need to know the details to do it - just use the AMC software!

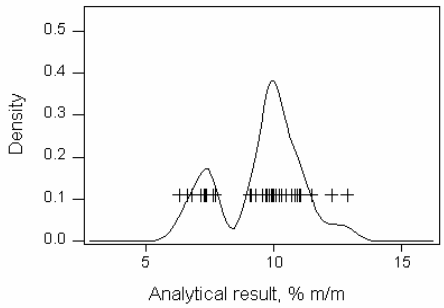

Figure 1. Results from a proficiency test for the determination of total sugars in condensed milk, showing the data (+) and a kernel density representation (line).

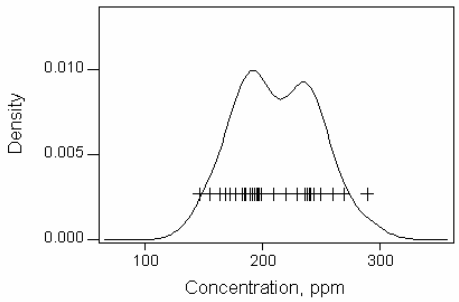

Figure 2. Concentration of aluminium in 40 flint artefacts from Southern England, showing data (+) and a kernel density representation (line).

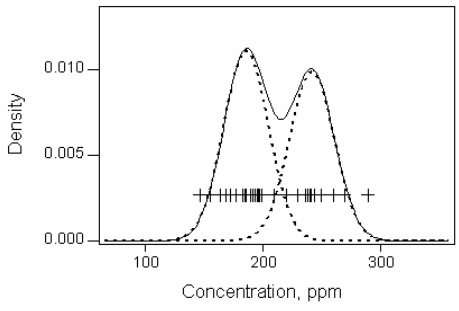

Figure 3. The flint data (+) from Figure 2 interpreted as a mixture (solid line) of two normally distributed components (dashed lines).

Useful reading

1. M Thompson. '*Using mixture models for bump-hunting in the results of proficiency tests.' Accred Qual Assur*, 2006 (in press).
2. M Aitkin, G T Wilson. '*Mixture models, outliers, and the EM algorithm.' Technometrics*, 1980, **22**, 325-331.

CPD Certification I certify that I have studied this document as a contribution to Continuing Professional Development.

Name…….....
Signature…….…….Date…….

Name of supervisor……..
Signature…….…….Date……

This Technical Brief was prepared for the Analytical Methods Committee by the Statistics Subcommittee (Chairman M Thompson), which is supported by the Food Standards Agency.

Maximum likelihood

Maximum likelihood is an estimation method that is more general than least squares. However, it often needs an iterative procedure to obtain the results. For data $(x_1,...,x_n)$ assumed independent and randomly drawn from a model defined by parameters $\boldsymbol{\theta}$, the likelihood is defined by $L(\boldsymbol{\theta}) = \prod_i f(x_i|\boldsymbol{\theta})$. We usually work with the log likelihood, $\log L(\boldsymbol{\theta}) = \sum_i \log f(x_i|\boldsymbol{\theta})$. The parameter estimates $\hat{\boldsymbol{\theta}}$ are those that maximise $\log L$. For a normal distribution we have $\log L = \sum_i \log(\exp(-(x_i-\mu)^2/2\sigma^2)/(\sqrt{2\pi}\sigma)$, where the parameters $\boldsymbol{\theta}$ are simply the mean μ and the variance σ^2. In that instance maximum likelihood gives the same outcome as the familiar least squares.

How the EM algorithm works

The EM (Expectation Maximisation) algorithm executes maximum likelihood estimation for mixture models. It is described here for a two-component normal mixture, where we want to estimate μ_1, σ_1, p_1, and μ_2, σ_2, p_2, for two components with proportions $p_1 + p_2 = 1$. Consider the data illustrated in Figure 4.

We start with initial guesses of the parameters $\hat{\mu}_1$, $\hat{\mu}_2$. In practice visual estimates are usually satisfactory, so we take the values at the modes, namely 14 and 18. Then we take the midpoint of the two means (16) to dichotomise the data. In terms of probabilities, we are saying that, as a first guess, data x_i falling below 16 belong to the first (lower) component with a probability $P_{1i} = 1$. Those falling above 16 belong to the first component with a probability $P_{1i} = 0$. We attribute complementary probabilities to the second component. These probabilities are shown in Figure 5.

Then the maximum likelihood estimates for components $j = 1, 2$ are:

$$\hat{p}_j = \sum_i P_{ji} / n$$

$$\hat{\mu}_j = \sum_i x_i P_{ji} / \sum_i P_{ji}$$

$$\hat{\sigma}^2 = \sum_j \sum_i ((x_i - \mu_j)^2 P_{ji}) / n$$

These are simply the formulae for means and a (pooled) variance, but with the data weighted by the probabilities. We can now calculate better estimates of the membership probabilities from

$$P_{ji} = \hat{p}_j f_j(x_i) / \sum_j \hat{p}_j f_j(x_i)$$

where $f_j(x_i)$ is the normal distribution density function, namely

$$f_j(x_i) = \exp(-(x_i - \hat{\mu}_j)^2 / 2\hat{\sigma}^2) / (\sqrt{2\pi}\,\hat{\sigma}).$$

The procedure is iterated until no worthwhile improvement in accuracy is obtained. For the example data, this gives the following parameter estimates:

$\hat{\mu}_1 = 13.9,\ \hat{p}_1 = 0.30,\ \hat{\sigma} = 1.1.$

$\hat{\mu}_2 = 17.7,\ \hat{p}_2 = 0.70.$

The final probabilities are shown in Figure 6, the mixture model in Figure 7.

AMC Health Warnings

1. It is easy to over-fit the data (that is, to use a model with too many components) because each extra component apparently improves the fit. Unfortunately, there is no simple and reliable method that tells us when to stop adding components, so we must rely on common sense.
2. It is usually safer to constrain the model components to a common variance (as in the example above). Allowing individual components to have different variances is easy, but might cause the algorithm to crash if outliers are present. Usually it is better to remove outliers before starting.

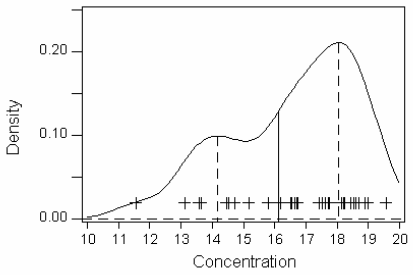

Figure 4. Example data (+) and the kernel density representation (curve), showing the position of the modes (dashed lines) and the midpoint (solid line).

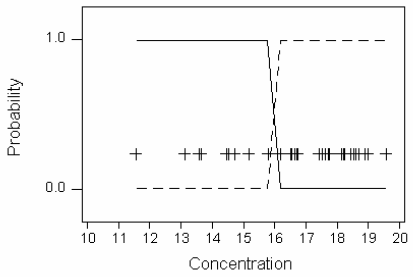

Figure 5. First guess of probabilities that individual data points (+) belong either to the lower component of the mixture model (P_1, solid line) or to the higher component (P_2, dashed line).

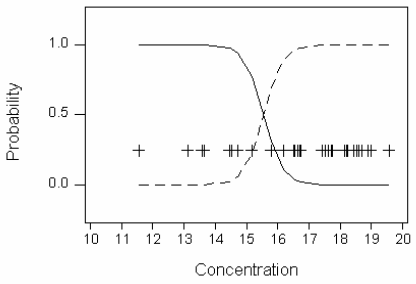

Figure 6. Probabilities that individual data points (+) belong either to the lower component of the mixture model (P_1, solid line) or to the higher component (P_2, dashed line), at the end of the iteration.

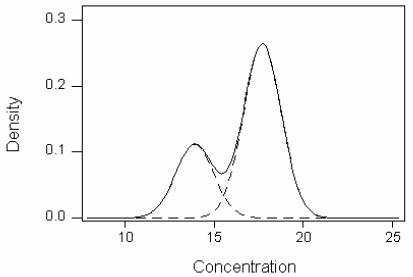

Figure 7. Data modelled as a mixture of two normal distributions (solid line). Individual components shown as dashed lines.

