New Bayesian methods for model comparison

How many components in a finite mixture?

Murray Aitkin

murray.aitkin@unimelb.edu.au

Department of Mathematics and Statistics
University of Melbourne
Australia
Bayesian model comparisons

Work supported by US National Center for Education Statistics and Australian Research Council.


Aim: to evaluate and develop general Bayesian model comparisons for arbitrary models through posterior likelihood ratios/posterior deviance differences.

Two aspects:

• Non-nested models – compute distribution of ratio of posterior likelihoods
• Nested models – compute posterior distribution of likelihood ratio
References for this work


Model comparisons – completely specified models

We have a random sample of data $\mathbf{y} = (y_1, \ldots, y_n)$ from a population which may be either of two completely specified distributions: Model 1: $f_1(y|\theta_1)$ and Model 2: $f_2(y|\theta_2)$, where $\theta_1$ and $\theta_2$ are known.

The likelihoods and priors are:

- $L_1 = \prod_i f_1(y_i|\theta_1)$ and $\pi_1$,
- $L_2 = \prod_i f_2(y_i|\theta_2)$ and $\pi_2 = 1 - \pi_1$.

Then by Bayes’ theorem,

$$
\frac{\pi_1|\mathbf{y}}{\pi_2|\mathbf{y}} = \frac{L_1}{L_2} \cdot \frac{\pi_1}{\pi_2}.
$$

For equal prior probabilities $\pi_1 = \pi_2$, the RHS is the likelihood ratio. So a likelihood ratio of 9 gives a posterior probability of $9/(9+1) = 0.9$ for Model 1.
Example – the Poisson-geometric choice (Cox 1961, 1962)

The data from Cox (1962) are \( n = 30 \) event counts \( y_i \) from either a Poisson or a geometric distribution, and are tabulated below as frequencies \( f \).

\[
\begin{array}{c|ccccc}
    y & 0 & 1 & 2 & 3 & > 3 \\
    \hline
    f & 12 & 11 & 6 & 1 & 0 \\
\end{array}
\]

We are given that the mean of the distribution is 0.8.
Likelihoods

The Poisson and geometric likelihoods and deviances (parametrised in terms of the means $\theta_1$ and $\theta_2$) are

$$L_P(\theta_1) = \prod_i e^{-\theta_1} \frac{\theta_1^{y_i}}{y_i!}$$

$$= e^{-n\theta_1} \frac{\theta_1^T}{F}$$

$$D_P(\theta_1) = -2 \log L_P(\theta_1) = 2[n\theta_1 - T \log \theta_1 + \log F]$$

$$L_G(\theta_2) = \prod_i \left( \frac{\theta_2}{1 + \theta_2} \right)^{y_i} \frac{1}{1 + \theta_2}$$

$$= \frac{\theta_2^T}{(1 + \theta_2)^{T+n}}$$

$$D_G(\theta_2) = -2 \log L_G(\theta_2) = 2[(T + n) \log (1 + \theta_2) - T \log \theta_2]$$

where $T = \sum_i y_i = 26$, $F = \prod_i y_i! = 384$. 
Likelihoods
Deviances
Likelihood ratio

At $\theta = 0.8$,

- $D_P(0.8) = 71.505$;
- $D_G(0.8) = 77.436$;
- $D_P(0.8) - D_G(0.8) = -5.931$;
- $L_P(0.8)/L_G(0.8) = e^{2.965} = 19.39$; with equal model priors,
- $\Pr[\text{Poisson} \mid \text{data, mean} = 0.8] = 19.39/20.39 = 0.951$.

We have very strong evidence in favour of the Poisson.

Now we are not give the mean.

How do we express the evidence for Poisson over geometric?

Common approaches – AIC, BIC – penalized maximized likelihoods; DIC – a penalized mean deviance.

The MLE of $\theta$ under both models is $\hat{\theta}_1 = \hat{\theta}_2 = \bar{y} = 26/30 = 0.867$. 
AIC, BIC

We need the maximized likelihoods under both models:

- $D_P(0.867) = 71.343$;
- $D_G(0.867) = 77.347$;
- $D_P(0.867) - D_G(0.867) = -6.004$.

Since both models have one parameter, the comparison of AICs and BICs is the same as that of the frequentist deviances $D_j(\hat{\theta}_j)$.

The Poisson is preferred.

The AIC, BIC and DIC comparisons are decision criteria: they do not give a measure of strength of preference: does the deviance difference of 6 give a strong or a weak preference for the Poisson?
Bayes factors

Bayes factors are commonly used to give the **weight of evidence**, but have well-known difficulties: improper priors on $\mu$ cannot be used.

Proper priors are essential, depending on hyper-parameters $\phi$.

Then the integrated likelihoods depend explicitly on the hyper-parameters $\phi$. 
Conjugate prior gamma \((r, \lambda)\)

\[
\bar{L}_P(r, \lambda) = \int L_P(\theta_1)\pi(\theta_1)d\theta_1
\]

\[
= \int \frac{1}{F}e^{-n\theta_1}\theta_1^T \cdot \frac{\lambda^r}{\Gamma(r)} e^{-\theta_1} \theta_1^{r-1}d\theta_1
\]

\[
= \frac{\Gamma(r + T)}{FT\Gamma(r)} \frac{\lambda^r}{(n + \lambda)^{r+T}}
\]

\[
= \frac{T!}{F} \cdot \binom{r + T - 1}{r - 1} \left(\frac{\lambda}{n + \lambda}\right)^r \left(\frac{n}{n + \lambda}\right)^T.
\]

We have gone, by integration over \(\theta_1\), from

- a Poisson likelihood with one parameter which we do not know to

- a scaled negative binomial likelihood with two parameters which we have to know.
We give the general approach, originally due to Dempster (1974, 1997).

The Poisson and geometric likelihoods are uncertain, because of our uncertainty about $\theta_1$ and $\theta_2$ in these models.

This uncertainty is expressed through the posterior distributions of $\theta_1$ and $\theta_2$, given the data and priors.

The Poisson likelihood $L_P(\theta_1)$ is a function of $\theta_1$, so we map the posterior distribution of $\theta_1$ into that of $L_P(\theta_1)$.

The geometric likelihood $L_G(\theta_2)$ is a function of $\theta_2$, so we map the posterior distribution of $\theta_2$ into that of $L_G(\theta_2)$.

This is very simply done by simulation, making random draws from the posteriors:
Posterior likelihoods

- make $M$ random draws $\theta_1^{[m]}$ from the gamma posterior distribution of $\theta_1$ under the Poisson model;
- substitute these draws into the Poisson likelihood, to give $M$ draws $L_P^{[m]} = L_P(\theta_1^{[m]})$ from the posterior distribution of the Poisson likelihood;
- make $M$ independent random draws $\theta_2^{[m]}$ from the posterior distribution of $\theta_2$ under the geometric model and prior;
- substitute these draws into the geometric likelihood, to give $M$ draws $L_G^{[m]} = L_G(\theta_2^{[m]})$ from the posterior distribution of the geometric likelihood;
- compute the $M$ ratios $L_P^{[m]} / L_G^{[m]}$: these are random draws from the posterior distribution of the likelihood ratio for Poisson to geometric.
Posterior deviances

We generally work with posterior deviances rather than posterior likelihoods, for reasons we show shortly – they are much better behaved.

We compute the two sets of posterior deviance draws:

- calculate $D_P^{[m]} = -2 \log L_P^{[m]}$, $D_G^{[m]} = -2 \log L_G^{[m]}$;

- compute the $M$ values of the deviance difference of Poisson to geometric by pairing the independent Poisson and geometric deviance draws:

$$DD_{PG}^{[m]} = D_P^{[m]} - D_G^{[m]};$$

- compute the $M$ values of the likelihood ratio of Poisson to geometric by exponentiating the deviance difference draws:

$$LR_{PG}^{[m]} = e^{-0.5 DD_{PG}^{[m]}};$$
Posterior deviances

- compute the $M$ values of the posterior probability of the Poisson, given equal prior probabilities (the indifference case):

$$\Pr^{[m]}[\text{Poisson} \mid \text{data}] = \frac{L^{[m]}_P}{L^{[m]}_P + L^{[m]}_G}.$$  

The $M$ values define the posterior distribution: we order them to give a picture of the cdf.
What prior?

Since we are working with the **posterior** of \( \theta \), the prior is less important: we are not **integrating** over the prior.

In particular, **we can work with flat or diffuse priors without any problem!**

For the Cox example, we use **flat priors on the means** \( \theta_1 \) and \( \theta_2 \): the posterior distribution of \( \theta_1 \) is \( \text{gamma}(T + 1, n) \), and that of \( \theta_2 \) is \( \text{beta}(T + 1, n - 1) \).

These give a **reference analysis**; this could be extended to **informative** priors if we wanted to use them.

We show the posterior **deviance distributions** for each model, and the posterior distribution of the **deviance difference**.
Posterior distributions of deviances

Poisson and geometric deviance distributions
Posterior distribution of deviance difference
Posterior distribution of Poisson posterior probability
Model preference

Of the 10,000 deviance differences, 96 are negative (geometric deviance smaller than Poisson deviance) – a proportion of 0.0096 (simulation SE 0.001).

The empirical posterior probability that the Poisson model fits better than the geometric (in likelihood) is 0.990 (SE 0.001).

This is equivalent to

• an empirical posterior probability of 0.990 (SE 0.001) that the posterior odds on the Poisson model, for unit prior odds, is greater than 1;

• an empirical posterior probability of 0.990 that the posterior probability of the Poisson model, with equal prior probabilities, is greater than 1/2.
The Poisson and geometric deviances are stochastically ordered –
the Poisson deviance distribution is stochastically smaller than – to the left of – the geometric distribution.

There is no crossing point of the cdfs (though they merge at $\infty$).

The deviance difference distribution is almost completely to the right of zero:

a random draw from the Poisson deviance distribution is almost always (empirical probability 0.990) smaller than a random draw from the geometric deviance distribution.
Posterior distribution of the posterior probability

The median deviance difference (Poisson-geometric) is \(-6.01\) – almost the same as the frequentist deviance difference – and the central 95% credible interval for the true deviance difference is \([-1.58, -10.64]\).

The median likelihood ratio (Poisson/geometric) is 20.0, and the 95% credible interval for the likelihood ratio is \([2.20, 204.4]\).

The median posterior probability of the Poisson model, given equal prior probabilities, is 0.953 (very close to the value 0.953 given that the mean is 0.8), and the 95% credible interval for it is \([0.647, 0.995]\).

The evidence in favour of the Poisson is quite strong, though not as strong as the ratio of maximized likelihoods suggests, because of the diffuseness of the posterior deviance difference distribution from the small sample.
Asymptotics

For regular models $f(y \mid \theta)$ with flat priors, giving an MLE $\hat{\theta}$ internal to the parameter space, the second-order Taylor expansion of the deviance $D(\theta) = -2\log L(\theta) = -2\ell(\theta)$ about $\hat{\theta}$ gives:

$$-2\ell(\theta) \doteq -2\ell(\hat{\theta}) - 2(\theta - \hat{\theta})' \ell'(\hat{\theta}) - (\theta - \hat{\theta})' \ell''(\hat{\theta})(\theta - \hat{\theta})$$

$$= -2\ell(\hat{\theta}) + (\theta - \hat{\theta})' I(\hat{\theta})(\theta - \hat{\theta})$$

$L(\theta) \doteq L(\hat{\theta}) \cdot \exp[-(\theta - \hat{\theta})' I(\hat{\theta})(\theta - \hat{\theta}) / 2]$}

$\pi(\theta \mid y) \doteq c \cdot \exp[-(\theta - \hat{\theta})' I(\hat{\theta})(\theta - \hat{\theta}) / 2]$
Asymptotic distributions

So asymptotically, given the data $y$, we have the posterior distributions:

$$\theta \sim N(\hat{\theta}, I(\hat{\theta})^{-1}),$$

$$(\theta - \hat{\theta})' I(\hat{\theta}) (\theta - \hat{\theta}) \sim \chi^2_p,$$

$$D(\theta) \sim D(\hat{\theta}) + \chi^2_p,$$

$$L(\theta) \sim L(\hat{\theta}) \cdot \exp(-\chi^2_p/2).$$

The likelihood $L(\theta)$ has a **scaled** $\exp(-\chi^2_p/2)$ distribution.

The deviance $D(\theta)$ has a **shifted** $\chi^2_p$ distribution, shifted by the frequentist deviance $D(\hat{\theta})$, where $p$ is the dimension of $\theta$.

The frequentist deviance is an **origin parameter** for the posterior deviance distribution: no random draw can give a smaller deviance value than the frequentist deviance.
Cox example

We extend the previous figure of the two deviance distributions with the corresponding asymptotic distributions:

- the asymptotic Poisson deviance distribution

\[
D_P(\theta_1) \sim D_P(\hat{\theta}_1) + \chi^2_1
= 71.343 + \chi^2_1,
\]

- the asymptotic geometric deviance distribution

\[
D_G(\theta_2) \sim D_G(\hat{\theta}_2) + \chi^2_1
= 77.347 + \chi^2_1.
\]

The empirical distributions are shown as solid curves, the asymptotic distributions are dashed curves. The agreement is very close for the Poisson, slightly worse for the geometric whose likelihood is more skewed.
Empirical and asymptotic deviance distributions
Model validation – the **saturated model**

So the evidence points strongly to the Poisson, if the Poisson and geometric are the only candidates.

But what about other models? – from a ML point of view, the “saturated" multinomial would always fit better!

We can easily extend the model comparison to three models, including the multinomial.

The multinomial likelihood and deviance, for counts $n_j$ at observed values $y_j$ with probabilities $p_j$, are

\[
L_M(\{p_j\}) = \prod_j p_j^{n_j},
\]

\[
D_M(\{p_j\}) = -2 \sum_j n_j \log p_j.
\]
Frequentist test

The frequentist deviance for the multinomial is 70.179.

The likelihood ratio test for the null Poisson to the alternative multinomial uses the frequentist deviance difference:

$$71.343 - 70.179 = 1.16.$$ 

This is compared with $$\chi^2_2$$.

The result is clearly **not significant at any conventional level**.

But is the asymptotic sampling distribution applicable? We don’t know.
Dirichlet prior and posterior

We use the conjugate Dirichlet prior:

$$\pi(\{p_j\}) = \frac{\Gamma(\sum_j a_j)}{\prod_j \Gamma(a_j)} p_j^{a_j-1},$$

giving the Dirichlet posterior

$$\pi(\{p_j\} \mid \{n_j\}) = \frac{\Gamma[\sum_j (a_j + n_j)]}{\prod_j \Gamma(a_j + n_j)} p_j^{a_j+n_j-1}.$$

For a non-informative analysis we take $a_j = 0 \forall j$, giving the posterior

$$\pi(\{p_j\} \mid \{n_j\}) = \frac{\Gamma(\sum_j n_j)}{\prod_j \Gamma(n_j)} p_j^{n_j-1}.$$
Deviance draws

• We make $M$ draws $p_j^{[m]}$ from the posterior,

• substitute them in the multinomial deviance to give

• $M$ multinomial deviance draws $D_M^{[m]} = D_M(\{p_j^{[m]}\})$,  

• order these

• and plot their empirical and asymptotic cdfs with those for the Poisson and geometric models.
Poisson, geometric and multinomial deviances

---

Bayesian Model Comparison – p. 32/74
Poisson-multinomial deviance
geometric-multinomial deviance
Model comparisons

Three major points:

• The agreement between empirical and asymptotic cdfs is not as close for the multinomial as for the parametric models:
  ◦ the heavier parametrization requires a larger sample size for asymptotic behaviour;
  ◦ the sample of 1 in the last category gives a highly skewed posterior in this parameter.

• Of the geometric-multinomial deviances, 605 are negative – an empirical proportion of 0.0605 – strong evidence against the geometric.

• Of the Poisson-multinomial deviances, 6154 are negative – an empirical proportion of 0.615 – we cannot choose clearly between the Poisson and the multinomial – there is no convincing preference for one over the other.
The galaxy recession velocity study – mixtures

• The data are the recession velocities of 82 galaxies from 6 well-separated sections of the Corona Borealis region.

• Do these velocities “clump” into groups or clusters, or does the velocity density increase initially and then gradually tail off?

  This has implications for theories of evolution of the universe.

• Investigate by fitting mixtures of normal distributions to the velocity data; the number of mixture components necessary to represent the data is the parameter of particular interest.
Recession velocities (/1000) of 82 galaxies

|    | 99 | 97 | 92 | 91 | 86 | 86 | 99 | 85 | 88 | 66 | 65 | 91 | 55 | 82 | 89 | 54 | 80 | 75 | 71 | 53 | 63 | 75 | 71 | 47 | 42 | 50 | 67 | 44 | 22 | 96 | 37 | 54 | 99 |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 78 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 56 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 48 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 35 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 17 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 08 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |

velocity
Data source

The data are given in


without the region, and are given with the region in:


We do not use the region – small samples in each region don’t provide much information.

Many analyses of these data; see

- McLachlan and Peel (1997)
Two normals
Three normals
Mixture of normals model

The general model for a $K$-component normal mixture has different means $\mu_k$ and variances $\sigma^2_k$ in each component:

$$f(y) = \sum_{k=1}^{K} \pi_k f(y|\mu_k, \sigma_k)$$

where

$$f(y|\mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k}} \exp \left\{ -\frac{1}{2\sigma_k^2} (y - \mu_k)^2 \right\}$$

and the $\pi_k$ are positive with $\sum_{k=1}^{K} \pi_k = 1$. 
Likelihood

Given a sample $y_1, ..., y_n$ from $f(y)$, the likelihood is

$$L(\theta) = \prod_{i=1}^{n} f(y_i)$$

and

$$= \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f(y_i | \mu_k, \sigma_k)$$

where $\theta = (\pi_1, ..., \pi_{K-1}, \mu_1, ..., \mu_K, \sigma_1, ..., \sigma_K)$.

The form of the likelihood is ugly and complicates likelihood analysis.
## ML results

<table>
<thead>
<tr>
<th>K</th>
<th>k</th>
<th>mean</th>
<th>prop</th>
<th>sd</th>
<th>mean</th>
<th>prop</th>
<th>sd</th>
<th>deviance</th>
<th>deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>20.83</td>
<td>1</td>
<td>4.54</td>
<td>20.83</td>
<td>1</td>
<td>4.54</td>
<td>480.83</td>
<td>480.83</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>21.49</td>
<td>0.533</td>
<td>4.49</td>
<td>21.35</td>
<td>0.740</td>
<td>1.88</td>
<td>480.83</td>
<td>440.72</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>20.08</td>
<td>0.467</td>
<td>19.36</td>
<td>0.260</td>
<td>8.15</td>
<td></td>
<td>480.83</td>
<td>440.72</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>32.94</td>
<td>0.037</td>
<td>2.08</td>
<td>33.04</td>
<td>0.037</td>
<td>0.92</td>
<td>425.36</td>
<td>406.96</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>21.40</td>
<td>0.877</td>
<td>21.40</td>
<td>0.878</td>
<td>2.20</td>
<td></td>
<td>425.36</td>
<td>406.96</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>9.75</td>
<td>0.086</td>
<td>9.71</td>
<td>0.085</td>
<td>0.42</td>
<td></td>
<td>425.36</td>
<td>406.96</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>33.04</td>
<td>0.037</td>
<td>1.32</td>
<td>33.05</td>
<td>0.037</td>
<td>0.92</td>
<td>416.50</td>
<td>395.43</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>23.50</td>
<td>0.352</td>
<td>23.50</td>
<td>0.665</td>
<td>2.27</td>
<td></td>
<td>416.50</td>
<td>395.43</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>20.00</td>
<td>0.526</td>
<td>19.75</td>
<td>0.213</td>
<td>0.45</td>
<td></td>
<td>416.50</td>
<td>395.43</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>9.71</td>
<td>0.085</td>
<td>9.71</td>
<td>0.085</td>
<td>0.42</td>
<td></td>
<td>416.50</td>
<td>395.43</td>
</tr>
</tbody>
</table>
## ML results

<table>
<thead>
<tr>
<th>K</th>
<th>k</th>
<th>mean</th>
<th>prop</th>
<th>sd</th>
<th>mean</th>
<th>prop</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1</td>
<td>33.04</td>
<td>0.037</td>
<td>1.07</td>
<td>33.05</td>
<td>0.036</td>
<td>0.92</td>
</tr>
<tr>
<td>2</td>
<td>26.38</td>
<td>0.037</td>
<td>22.92</td>
<td>0.289</td>
<td>23.04</td>
<td>0.289</td>
<td>1.02</td>
</tr>
<tr>
<td>3</td>
<td>23.04</td>
<td>0.366</td>
<td>21.85</td>
<td>0.245</td>
<td>22.93</td>
<td>0.424</td>
<td>3.05</td>
</tr>
<tr>
<td>4</td>
<td>19.76</td>
<td>0.475</td>
<td>19.82</td>
<td>0.344</td>
<td>19.93</td>
<td>0.453</td>
<td>0.63</td>
</tr>
<tr>
<td>5</td>
<td>9.71</td>
<td>0.085</td>
<td>9.71</td>
<td>0.085</td>
<td>9.71</td>
<td>0.085</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>410.85</td>
<td></td>
<td></td>
<td>392.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>33.04</td>
<td>0.037</td>
<td>0.81</td>
<td>33.04</td>
<td>0.037</td>
<td>0.92</td>
</tr>
<tr>
<td>2</td>
<td>26.24</td>
<td>0.044</td>
<td>26.98</td>
<td>0.024</td>
<td>26.98</td>
<td>0.024</td>
<td>0.018</td>
</tr>
<tr>
<td>3</td>
<td>23.05</td>
<td>0.357</td>
<td>22.93</td>
<td>0.424</td>
<td>22.93</td>
<td>0.424</td>
<td>1.20</td>
</tr>
<tr>
<td>4</td>
<td>19.93</td>
<td>0.453</td>
<td>19.79</td>
<td>0.406</td>
<td>19.79</td>
<td>0.406</td>
<td>0.68</td>
</tr>
<tr>
<td>5</td>
<td>16.14</td>
<td>0.025</td>
<td>16.13</td>
<td>0.024</td>
<td>16.13</td>
<td>0.024</td>
<td>0.043</td>
</tr>
<tr>
<td>6</td>
<td>9.71</td>
<td>0.085</td>
<td>9.71</td>
<td>0.085</td>
<td>9.71</td>
<td>0.085</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>394.58</td>
<td></td>
<td></td>
<td>365.15</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### ML results

<table>
<thead>
<tr>
<th>K</th>
<th>k</th>
<th>mean prop sd</th>
<th>mean prop sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>1</td>
<td>33.04 0.037 0.66</td>
<td>33.04 0.037 0.92</td>
</tr>
<tr>
<td>2</td>
<td>26.60 0.033</td>
<td>26.98 0.024 0.018</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>23.88 0.172</td>
<td>23.42 0.300 0.99</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>22.31 0.221</td>
<td>22.13 0.085 0.25</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>19.83 0.427</td>
<td>19.89 0.444 0.73</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>16.13 0.024</td>
<td>16.13 0.024 0.043</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>9.71 0.085</td>
<td>9.71 0.085 0.42</td>
<td></td>
</tr>
<tr>
<td></td>
<td>388.87</td>
<td>363.04</td>
<td></td>
</tr>
</tbody>
</table>

Bayesian Model Comparison – p. 46/74
Three normals
Evidence for components

- For all models with $K < 9$, the smallest 7 and largest 3 observations form stable components. Note that $7/82 \approx 0.085$, $3/82 \approx 0.037$.

- For $K = 4$ the central group of observations is split into two smaller groups with means around 22 and 20.

- Further increases in $K < 9$ fragment these two subgroups into smaller subgroups and shards.

The evidence for three or four components looks quite strong.
Bayes analysis

Analysis is greatly simplified by the introduction of a set of latent Bernoulli variables $z_k$ for membership in component $k$; this allows the “complete data" representation

$$f^*(k, \pi, z, \theta, y) = f(k) f(\pi|k) f(z|\pi, k) f(\theta|z) f(y|\theta, z),$$

where $\theta$ is the set of mean and standard deviation parameters $\mu_k, \sigma_k$.

This allows simpler conditional distributions in the MCMC algorithm, as in the EM algorithm.

Many Bayesian analyses use an additional layer of prior structure, in which the priors for $k, \pi$ and $\theta$ depend on hyperparameters which have to be specified.
Bayes analysis

- All the Bayes analyses use some form of Data Augmentation or Markov chain Monte Carlo analysis, with
  - updating of the successive conditional distributions of each set of parameters and the latent component membership variables
  - given the others and the data $y$.

- Most of the analyses
  - take $K$ initially as fixed
  - obtain an integrated likelihood over the other parameters for each $K$, and
  - use Bayes’ theorem to obtain the posterior probabilities of each value of $K$. 

Bayesian Model Comparison  – p. 50/74
Bayes analysis

More complex analyses (Richardson and Green 1997) use Reversible Jump MCMC in which $K$ is included directly in the parameter space, which changes as $K$ changes, as jumps are allowed across different values of $K$.

R & G also gave posterior deviance distributions for each number of components (for a different data set) as kernel densities,

which overlapped considerably,

making it difficult to identify an appropriate number of components.
Priors

The choice of prior distributions has varied among Bayes analysts for the analyses of the galaxy data by

- Escobar and West (1995)
- Carlin and Chib (1995)
- Phillips and Smith (1996)
- Roeder and Wasserman (1997) and
- Richardson and Green (1997).
## Prior distributions for $K$

<table>
<thead>
<tr>
<th>$K$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>EW</td>
<td>.01</td>
<td>.06</td>
<td>.14</td>
<td>.21</td>
<td>.21</td>
<td>.17</td>
<td>.11</td>
<td>.06</td>
<td>.02</td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>-</td>
<td>.33</td>
<td>.33</td>
<td>.33</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PS</td>
<td>.16</td>
<td>.24</td>
<td>.24</td>
<td>.18</td>
<td>.10</td>
<td>.05</td>
<td>.02</td>
<td>.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RW</td>
<td>.10</td>
<td>.10</td>
<td>.10</td>
<td>.10</td>
<td>.10</td>
<td>.10</td>
<td>.10</td>
<td>.10</td>
<td>.10</td>
<td>.10</td>
</tr>
</tbody>
</table>
| RG  | .03| .03| .03| .03| .03| .03| .03| .03| .03| .03|...
### Posterior distributions for $K$

<table>
<thead>
<tr>
<th>$K$</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>EW1</td>
<td>.03</td>
<td>.11</td>
<td>.22</td>
<td>.26</td>
<td>.20</td>
<td>.11</td>
<td>.05</td>
<td>.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EW2</td>
<td>.02</td>
<td>.05</td>
<td>.14</td>
<td>.21</td>
<td>.21</td>
<td>.16</td>
<td>.11</td>
<td>.06</td>
<td>.03</td>
<td>.01</td>
<td></td>
</tr>
<tr>
<td>CC1</td>
<td>.64</td>
<td>.36</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CC2</td>
<td>.004</td>
<td>.996</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PS</td>
<td></td>
<td></td>
<td>.03</td>
<td>.39</td>
<td>.32</td>
<td>.22</td>
<td>.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RW</td>
<td>.999</td>
<td>.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RG</td>
<td>.06</td>
<td>.13</td>
<td>.18</td>
<td>.20</td>
<td>.16</td>
<td>.11</td>
<td>.07</td>
<td>.04</td>
<td>.02</td>
<td>.01</td>
<td>.01</td>
</tr>
<tr>
<td>N</td>
<td>.02</td>
<td>.13</td>
<td>.16</td>
<td>.25</td>
<td>.20</td>
<td>.13</td>
<td>.06</td>
<td>.03</td>
<td>.01</td>
<td>.01</td>
<td></td>
</tr>
</tbody>
</table>
Posterior deviance distributions

For each $K$ we report the analysis from Celeux, Forbes, Robert and Titterington ("Deviance information criteria for missing data models" – Bayesian Analysis) using a diffuse Dirichlet prior on the component proportions $\pi_k$ and diffuse conjugate priors on the means $\mu_k$ and inverse variances $\sigma_k^2$.

After convergence 10,000 values were sampled from this posterior distribution (at long intervals in the sequence of draws to minimise autocorrelation),

and the $K$-component mixture likelihood computed for each parameter set.

The figures show the deviance distributions for $K = 1, 2, \ldots, 7$ (solid), with the asymptotic $\chi^2_{3K-1}$ distribution (dotted).
1 component
2 components
3 components
4 components
6 components
7 components
Asymptotic behaviour

The deviance draw distribution agrees almost exactly with its asymptotic form \( \chi^2_2 \) for \( K = 1 \).

As \( K \) increases, two different phenomena are visible:

- the frequentist deviance moves further away from the deviance draw distribution – the frequentist deviance is increasingly unrepresentative of the deviance draw distribution minimum

- the deviance draw distribution is increasingly more diffuse than the asymptotic distribution – the data are spread increasingly thinly over the increasing number of parameters, so all posteriors are more diffuse, as is the deviance posterior.

We now show the deviance distributions on the same scale.
1-7 components

![CDF graph with deviance on the x-axis and CDF on the y-axis, showing curves for 1 to 7 components.](image-url)
How many components?

- The deviance distribution for $K = 2$ greatly improves on the single normal.
- The improvement continues for $K = 3$.
- As the number of components increases beyond 3 the deviance distributions move steadily to the right, to larger values (lower likelihoods).
- They also become more diffuse, with increasing slope.
- It is clear that components beyond 3 add only noise.

The number of negative deviance differences, for 2,4,6,7 compared to 3, are for 2: 9620; for 4: 6098; for 5: 6921; for 6: 6591; for 7: 8226.
Deviance difference 3-2 components

Bayesian Model Comparison – p. 66/74
Deviance difference 3-4 components

Bayesian Model Comparison – p. 67/74
Deviance difference 3-5 components
Deviance difference 3-6 components

dev3-dev6
cdf
Deviance difference 3-7 components
Comparisons

• Comparing the deviances for 3 and 4 components, the difference is centred around 2.5, in favour of the 3-component model.

• This is not compelling evidence for 3 rather than 4, but it is much less compelling evidence for 4 rather than 3.

• This is the standard issue of parsimony – there is no compelling evidence for more than 3 components, so we choose 3 as the parsimonious representation of the mixture.

• Celeux et al come to the same conclusion - the DIC penalties they examine all favour the 3-component model.

• The likelihood distribution approach does not require any penalty because we do not use a single number to represent the model likelihood, but its full distribution.
Posterior probabilities from median deviances

If we want to calculate posterior model probabilities as single numbers, it would be more reasonable to use posterior median likelihoods than integrated likelihoods.

The median deviances, and corresponding posterior probabilities, are given in the table.

Table 1: Posterior median deviances and model probabilities

<table>
<thead>
<tr>
<th>K</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev</td>
<td>482.2</td>
<td>448.9</td>
<td>431.0</td>
<td>433.9</td>
<td>436.2</td>
<td>438.3</td>
<td>440.8</td>
</tr>
<tr>
<td>prob</td>
<td>0</td>
<td>0</td>
<td>0.745</td>
<td>0.175</td>
<td>0.055</td>
<td>0.019</td>
<td>0.006</td>
</tr>
</tbody>
</table>
Read more soon!


Watch for it!