New Bayesian methods for model comparison in astrostatistics

How many components in a finite mixture?

Murray Aitkin

murray.aitkin@unimelb.edu.au

School of Mathematics and Statistics

University of Melbourne

Australia

Acknowledgements

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

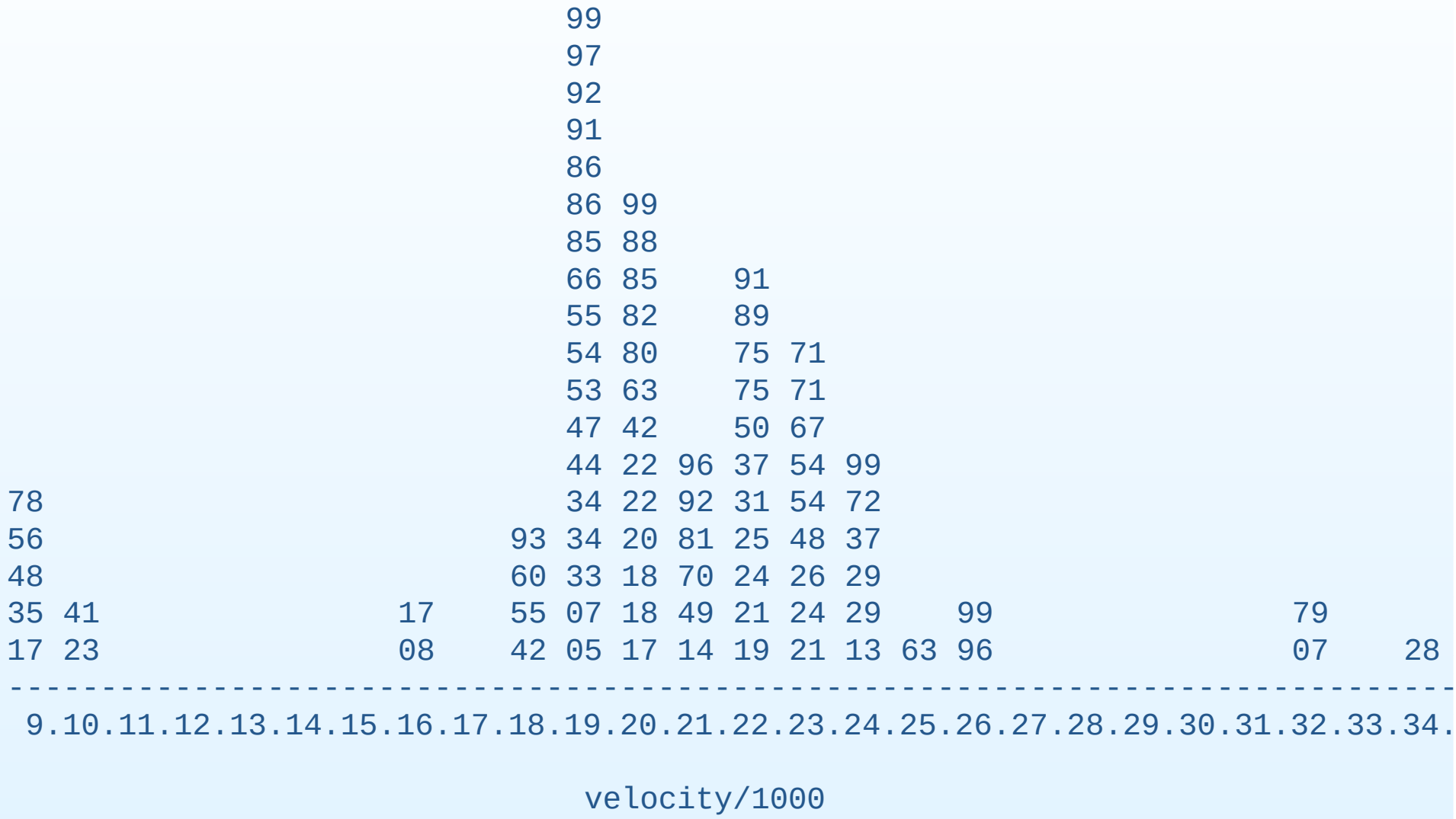
Book treatment:
Murray Aitkin (2010) *Statistical Inference: an Integrated Bayesian/Likelihood Approach*. Chapman and Hall/CRC

The galaxy recession velocity study

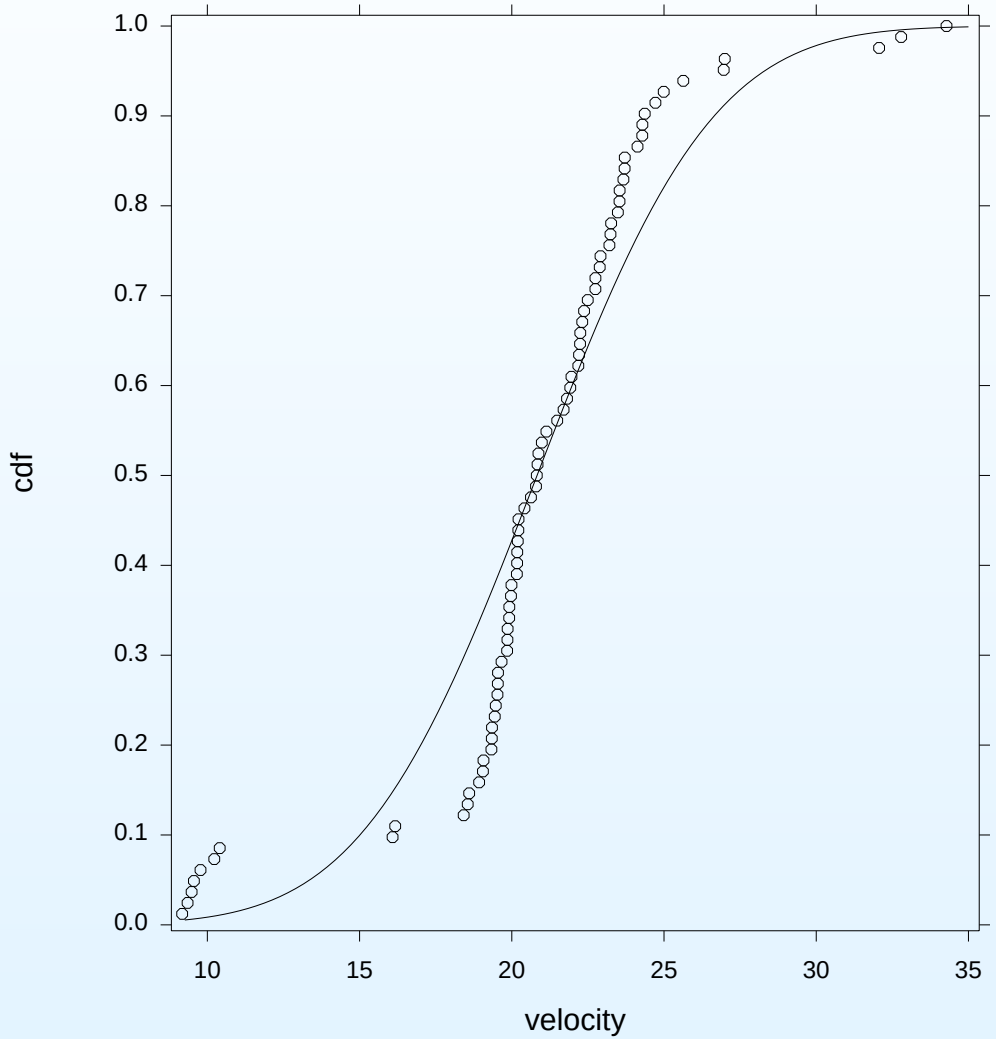
- The data are the recession velocities of 82 galaxies from 6 well-separated sections of the Corona Borealis region (Postman, Huchra and Geller, The Astronomical Journal 1986).
- Do these velocities clump into groups or clusters, or does the velocity density increase initially and then gradually tail off?

  This had implications for theories of evolution of the universe. If the velocities clump, the velocity distribution should be multi-modal.
- Investigated by fitting mixtures of normal distributions to the velocity data; the number of mixture components necessary to represent the data – or the number of modes – is the parameter of particular interest.

Recession velocities in km/sec (/1000) of 82 galaxies



Could these come from a single normal distribution?



Clearly not!

Clumping by mixtures

Mixture distributions are widely used to represent heterogeneity; mixtures of normals are the most common for data on a continuous scale.

The general model for a K-component normal mixture has different means μ_k and variances σ_k^2 in component k, which makes up a proportion π_k of the population:

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

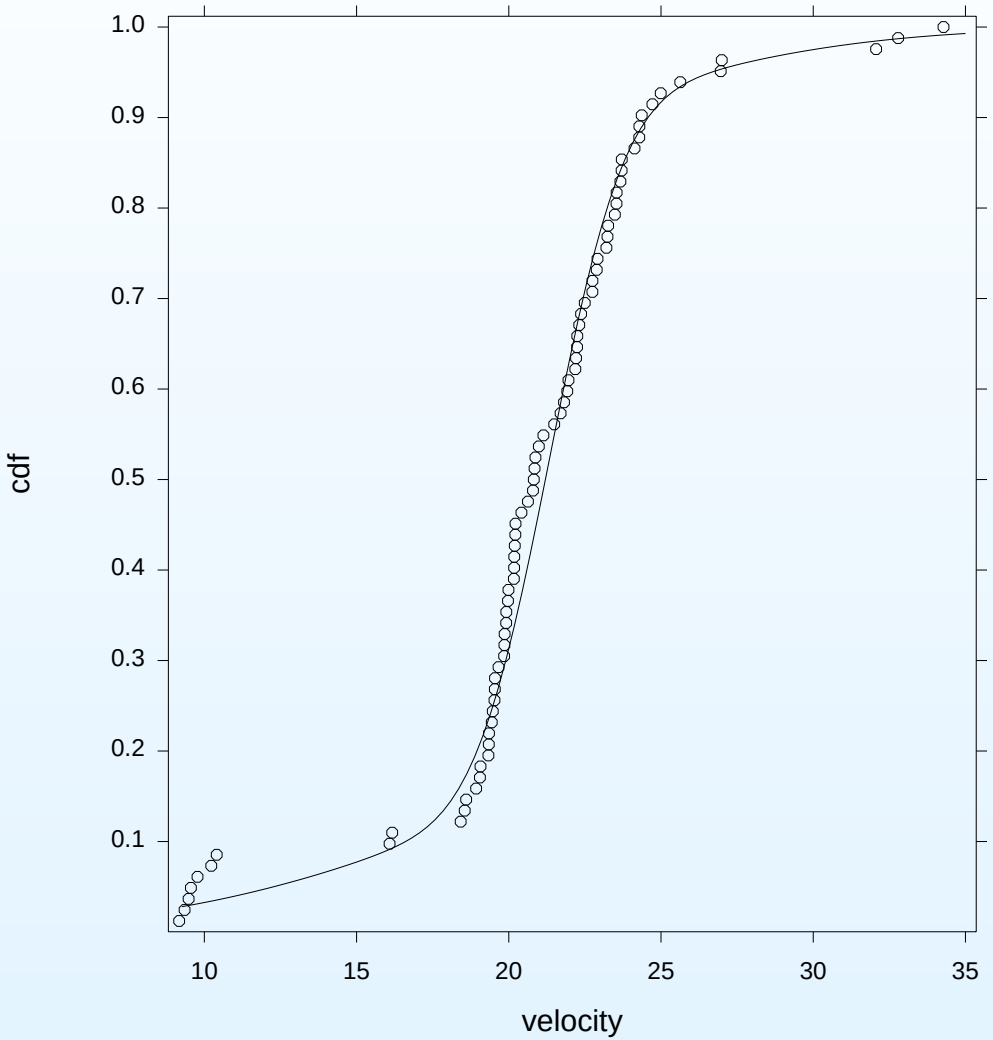
where

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

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

Fitting by ML is straightforward with an EM algorithm.

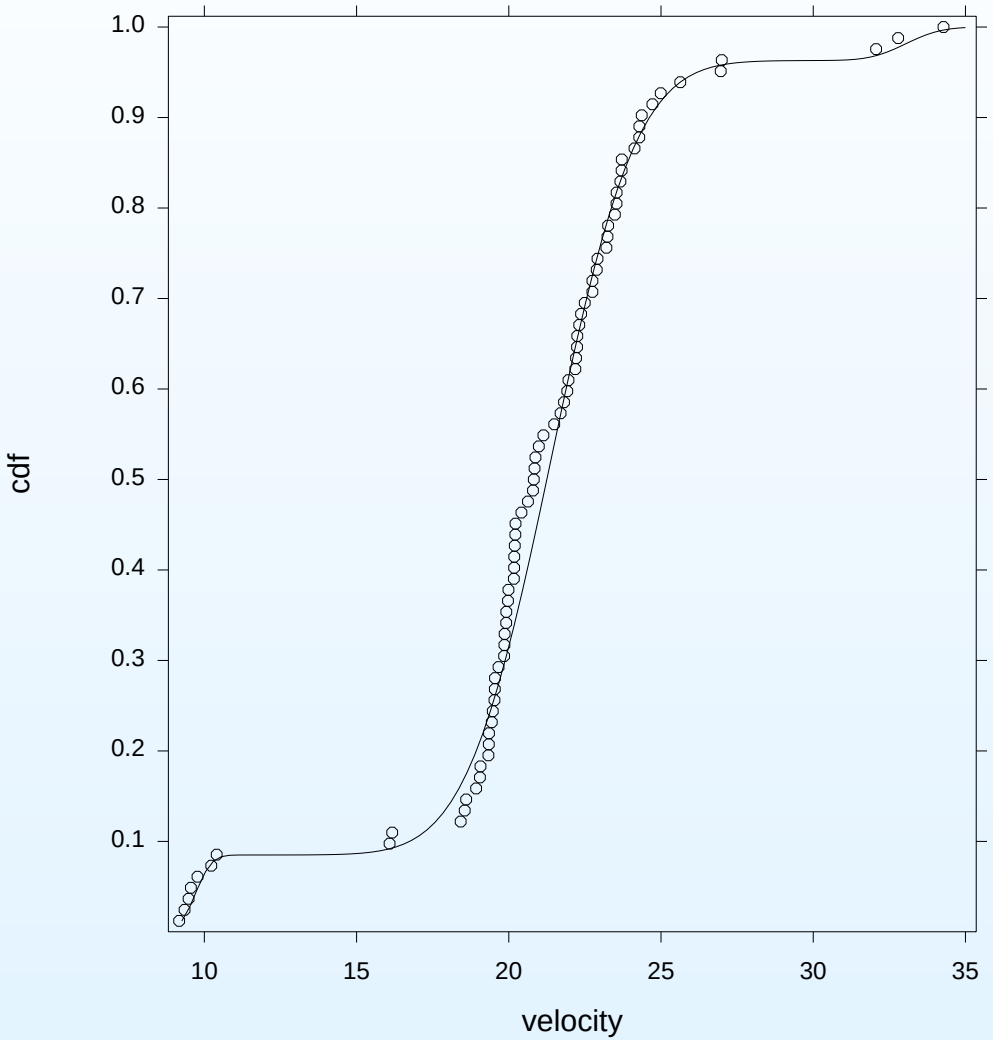
Two normals



$$\hat{\mu}_1 = 21.35, \hat{\sigma}_1 = 1.88, \hat{\pi}_1 = 0.740$$

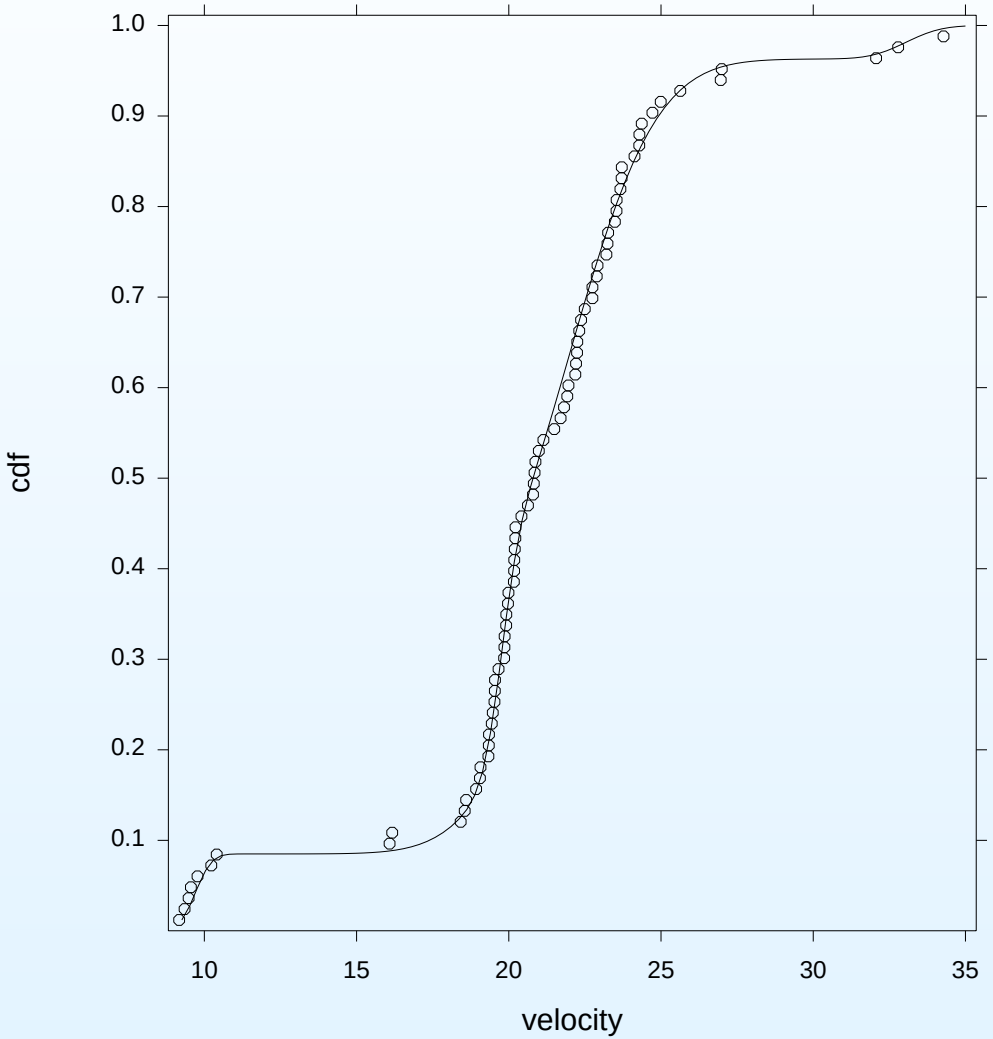
$$\hat{\mu}_2 = 19.36, \hat{\sigma}_2 = 8.15, \hat{\pi}_2 = 0.260$$

Three normals



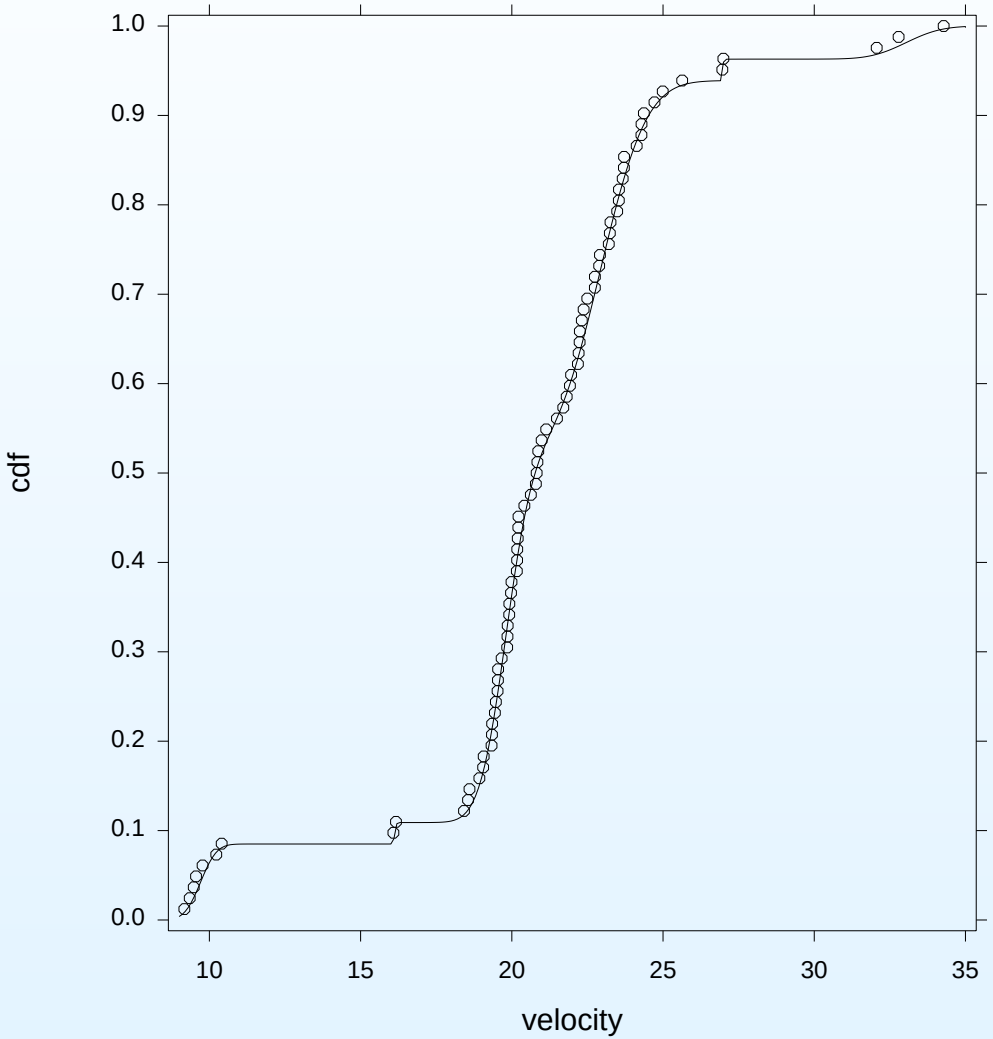
A good fit?

Four normals



A very close fit!

Six normals



Is this really a better fit?

Roeder's analysis

Roeder computed a nonparametric estimate of the velocity density by optimizing a goodness-of-fit criterion based on spacings between the ordered observations.

A simple kernel density estimate suggested three or more components; histograms gave inconsistent impressions depending on the bandwidth and limits.

A surface generated by bandwidth variations suggested three or more components.

Roeder kernel density graph

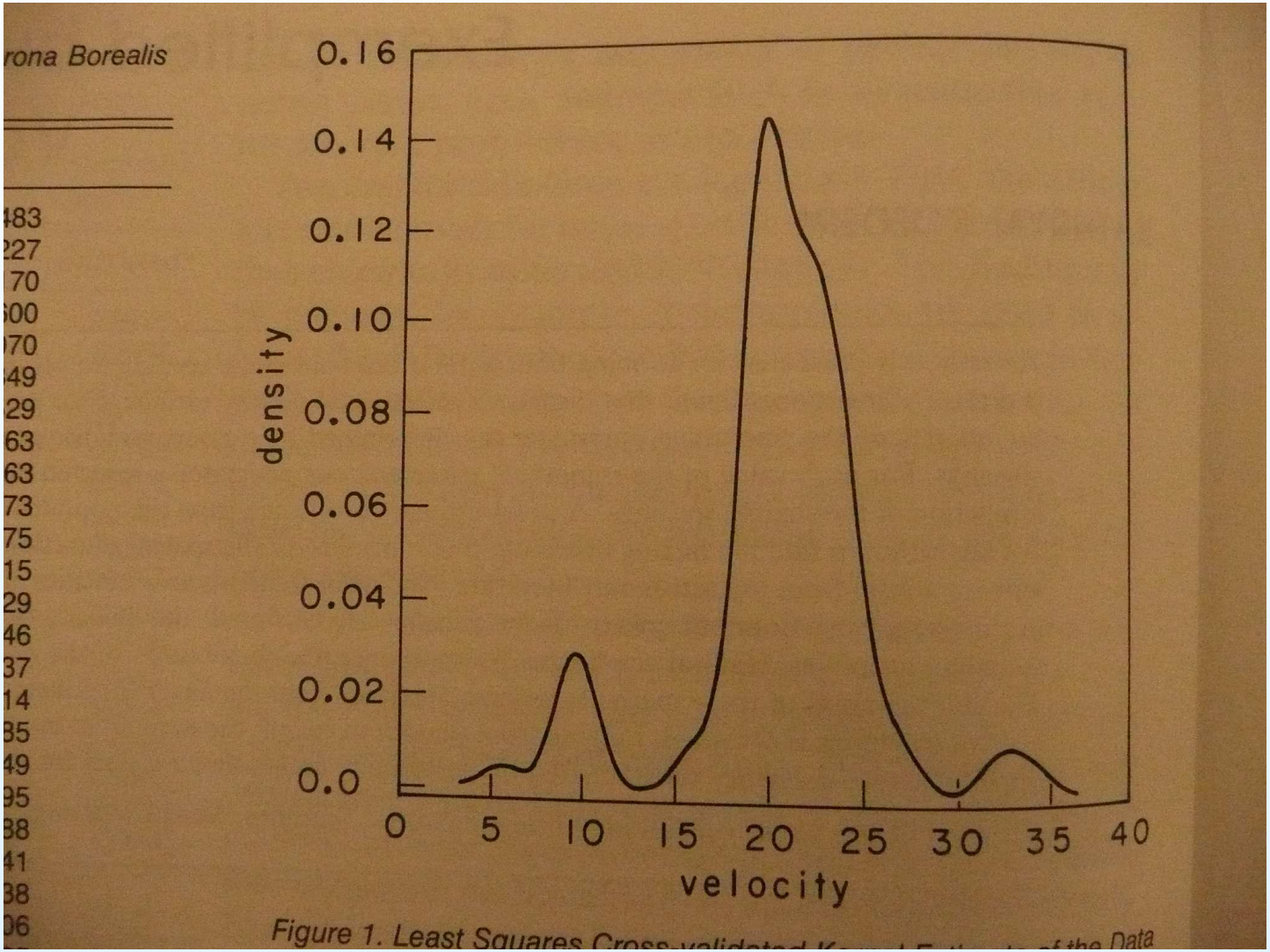


Figure 1. Least Squares Cross-validated

Roeder histograms graph

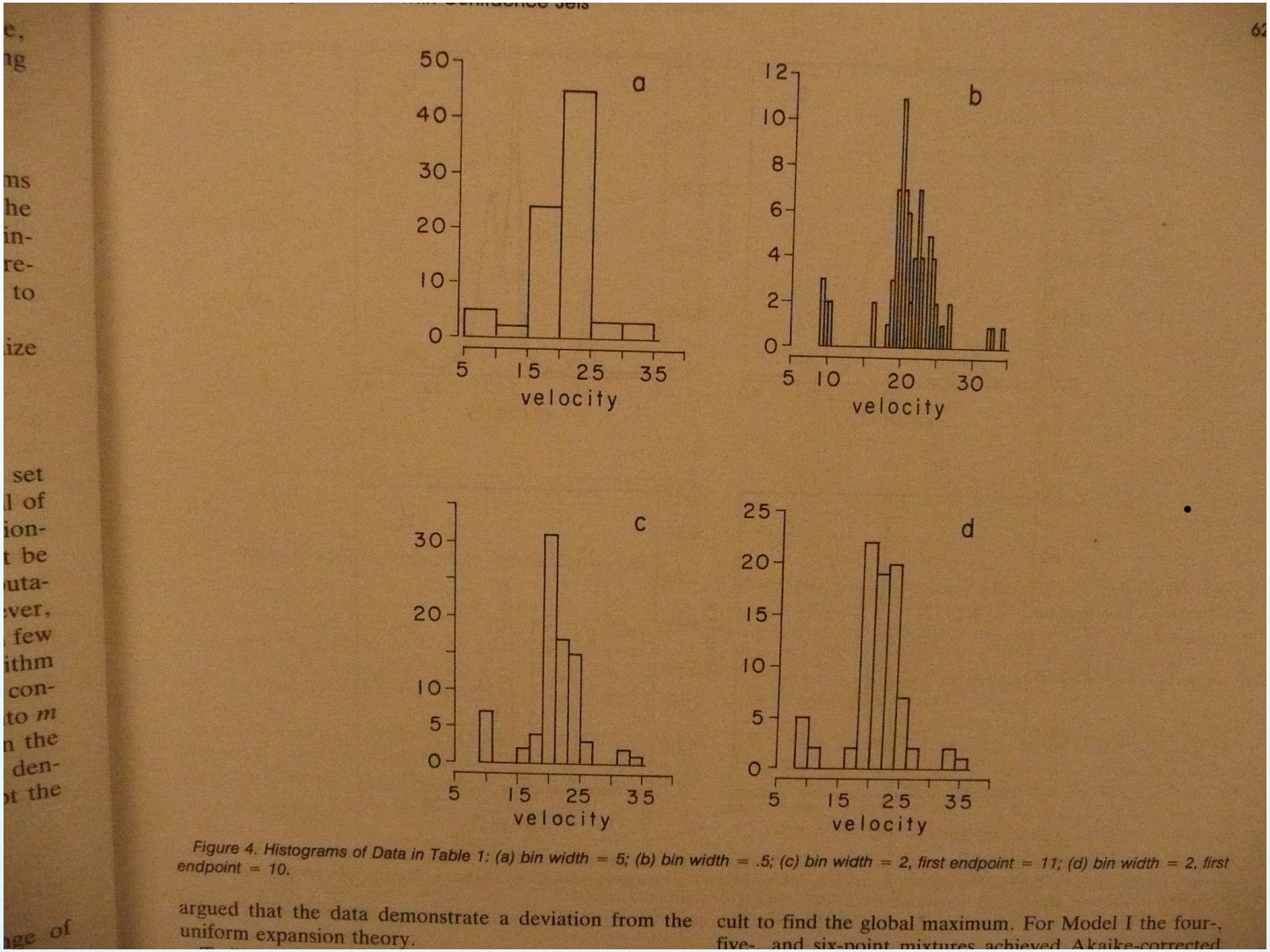


Figure 4. Histograms of Data in Table 1: (a) bin width = 5; (b) bin width = .5; (c) bin width = 2, first endpoint = 11; (d) bin width = 2, first endpoint = 10.

argued that the data demonstrate a deviation from the uniform expansion theory.

cult to find the global maximum. For Model I the four-, five- and six-point mixtures achieved Akaike-corrected

Roeder density surface graph

(1982), and Silverman
od of density estimation
e smoothing parameter.
hat has produced good
Rudemo 1982). Non-
at use cross-validation
meter have the advan-
l known that the point-
is large—$O(n^{-2/3})$ for
el estimators—LSCV
densities are plausible

n density by a normal

$$\left(\frac{\theta}{}\right) dQ(\theta), \qquad (1)$$

nsity, h is the smooth-
y probability measure
is an extremely rich
restricted, any density
normal mixture. Pre-

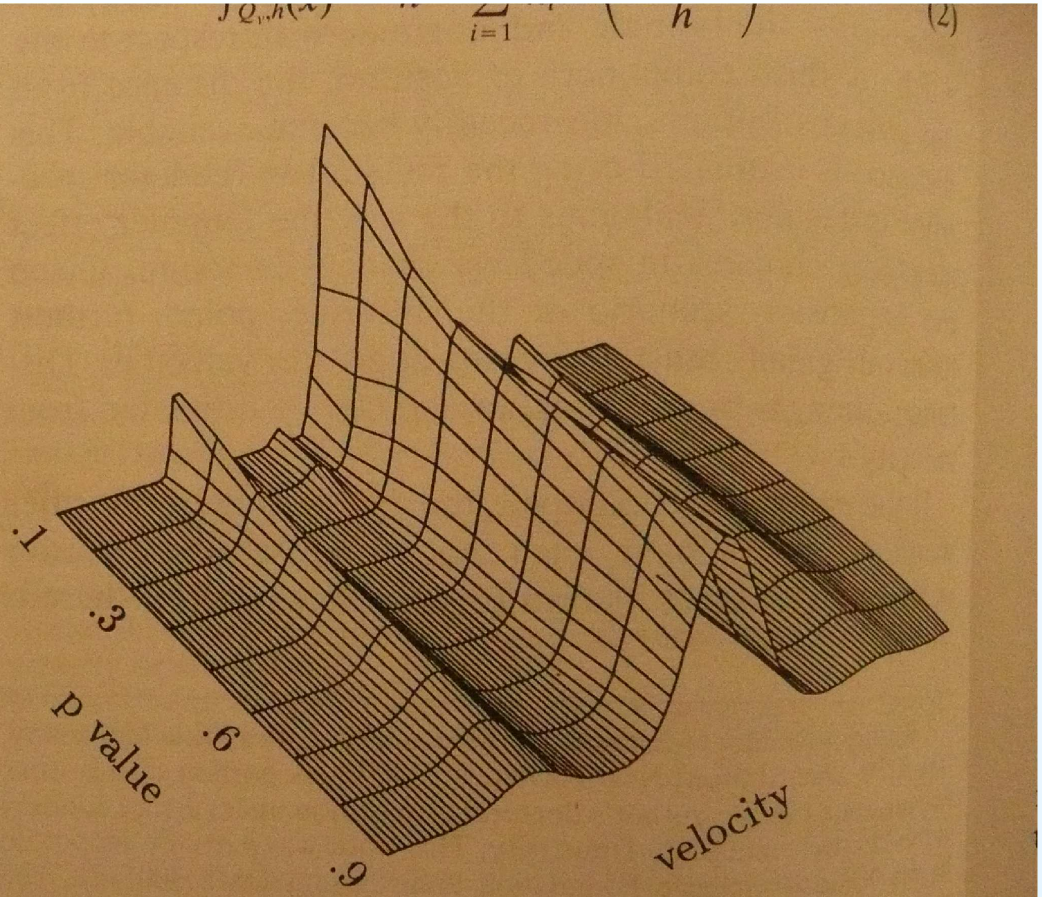


Figure 2. Confidence Set of Density Estimates. The foreground is the estimate associated with a p value of .10, and the background is the estimate associated with a p value of .90.

Model assessment and comparison

How do we decide whether four components are needed, more than four, or less than four?

The most common frequentist methods are AIC, BIC and the bootstrap likelihood ratio test.

These don't work well (see references for explanation).

The most common Bayesian method uses the integrated likelihood.

This doesn't work well either.

The DIC is used but is widely criticised.

We consider first the simplest case of completely specified models.

Model comparisons – completely specified models

We have a random sample of data $\mathbf{y} = (y_1, ..., y_n)$ 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 θ_1 and θ_2 are known.

The likelihoods and priors for the two models are:

- $L_1 = \prod_i f_1(y_i|\theta_1)$ and π_1,
- $L_2 = \prod_i f_2(y_i|\theta_2)$ and $\pi_2 = 1 - \pi_1$.

Then by Bayes's theorem, the posterior odds (ratio of posterior probabilities) for model 1 over model 2 is

$$\frac{\pi_{1|\mathbf{y}}}{\pi_{2|\mathbf{y}}} = \frac{L_1}{L_2} \cdot \frac{\pi_1}{\pi_2}.$$

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.

Model comparisons – general models

The models are Model 1 – $f_1(y|\theta_1)$ and Model 2 – $f_2(y|\theta_2)$ as before, but now θ_1 and θ_2 are unspecified except for priors $\pi_1(\theta_1)$ and $\pi_2(\theta_2)$.

The likelihood ratio now depends on θ_1 and θ_2.

We eliminate this dependence by integrating out the unknown parameters with respect to their priors, to give integrated likelihoods:

$\bar{L_1} = \int L_1(\theta_1)\pi_1(\theta_1)\mathrm{d}\theta_1$, $\bar{L_2} = \int L_2(\theta_2)\pi_2(\theta_2)\mathrm{d}\theta_2$.

The Bayes factor (for the relative support for Model 1 over Model 2) is defined to be $\bar{L_1}/\bar{L_2}$, and the posterior odds on Model 1 over Model 2 is defined to be

$$\frac{\bar{L_1}}{\bar{L_2}} \cdot \frac{\pi_1}{\pi_2},$$

as though the integrated likelihoods are from completely specified models.

Integrated likelihood difficulties

This approach to model comparisons is not restricted to two models: it can handle any number of competing models in the same way.

However, it has well-known difficulties:

- Improper priors cannot be used, as the integration leaves an arbitrary constant in the integrated likelihood.
- Proper priors are informative, depending on prior parameters ϕ.
- Then the integrated likelihood $\bar{L}(\phi)$ depends explicitly on the prior parameter ϕ –
- A change in the value of the prior parameter will change the value of the integrated likelihood.

Bayes analysis of mixtures

Bayesian and ML analyses are greatly simplified by the introduction of a set of latent Bernoulli variables $\{Z_{ik}\}$ for membership of observation i in component k.

This allows the complete data representation (with the Z_{ik} counterfactually observed)

$$f^*(y_i, \{Z_{ik}\}) = \prod_{k=1}^{K} f(y_i|\mu_k, \sigma_k)^{Z_{ik}} \cdot \pi_k^{Z_{ik}}$$

We can then write the complete data likelihood as

$$L^*(Z, \theta, \pi, k) = \prod_{i=1}^{n} \prod_{k=1}^{K} f(y_i|\mu_k, \sigma_k)^{Z_{ik}} \cdot \pi_k^{Z_{ik}}.$$

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

Posterior model probabilities for the galaxy data

- All the Bayes analyses used some form of Data Augmentation or Markov chain Monte Carlo analysis, with
  - updating of the successive conditional distributions of the set of parameters and
  - the set of latent component membership variables
  - given the other set and the data y.
- Most of the analyses
  - took K initially as fixed,
  - obtained an integrated likelihood over the other parameters for each K depending on the settings of the prior parameters,
  - and used Bayes's theorem to obtain the posterior probabilities of each value of K.

Bayes analysis

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

The choice of prior distributions (including for K) varied among Bayes analyses of the galaxy data by

- Escobar and West (1995)
- Carlin and Chib (1995)
- Phillips and Smith (1996)
- Roeder and Wasserman (1997)
- Richardson and Green (1997) and
- Nobile (2004).

Prior distributions for K

K	1	2	3	4	5	6	7	8	9	10
EW	.01	.06	.14	.21	.21	.17	.11	.06	.02	
CC	-	.33	.33	.33	-	-	-	-	-	
PS	.16	.24	.24	.18	.10	.05	.02	.01		
RW	.10	.10	.10	.10	.10	.10	.10	.10	.10	.10
RG	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03...
N	?									

Posterior distributions for K

K	3	4	5	6	7	8	9	10	11	12	13
EW1		.03	.11	.22	.26	.20	.11	.05	.02		
EW2	.02	.05	.14	.21	.21	.16	.11	.06	.03	.01	
CC1	.64	.36	-	-	-	-	-	-	-	-	-
CC2	.004	.996	-	-	-	-	-	-	-	-	-
PS				.03	.39	.32	.22	.04			
RW	.999	.00									
RG	.06	.13	.18	.20	.16	.11	.07	.04	.02	.01	.01
N	.02	.13	.16	.25	.20	.13	.06	.03	.01	.01	

Conclusions?

Most posteriors were diffuse, with modes at 6 or 7 components.

Carlin and Chib found 3 or 4 depending on their priors.

Roeder and Wasserman found 3 with probability almost 1.

So how many components are there?

“Explanation: the results are different because of the different priors."

This is not the solution: it is the problem.

Solution: posterior likelihoods/deviances

We give the general approach, originally due to Dempster.

The model likelihoods are uncertain, because of our uncertainty about the parameters in these models.

The parameter uncertainty is expressed through the posterior distributions of each model's parameters θ_k, given the data and priors.

The model k likelihood $L_k(\theta_k)$ is a functional – a function of both θ_k and the observed data, so we map the posterior distribution of θ_k into that of $L_k(\theta_k)$.

This is very simply done by simulation, making random draws from the posteriors and substituting them in the likelihoods.

Non-informative priors and MCMC for finite mixtures

For each $K = 1, 2, ...$ we use a diffuse Dirichlet prior on the component proportions π_k and diffuse conjugate priors on the means μ_k and inverse variances $1/\sigma_k^2$, for $k = 1, 2, ...K$.

For computational MCMC analysis these have to be proper, so slightly informative.

- The MCMC sampler is run till convergence of the joint posterior distribution of the parameter set for each K.
- Then we sample $M = 10,000$ values $\theta_k^{[m]}$ for each component from this posterior distribution, and
- compute the K-component mixture likelihood $L_K^{[m]} = L_K(\theta_1^{[m]}, ..., \theta_K^{[m]})$ for each parameter set.

This study was done for the galaxy data by Celeux et al (2006), in an evaluation of various rules for penalizing the posterior mean deviance in the DIC of Spiegelhalter et al (2002).

Asymptotics for likelihoods and deviances

We generally work with posterior deviances rather than posterior likelihoods – their asymptotics are much better behaved.

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:

$$
\begin{aligned}
-2\ell(\theta) &\doteq -2\ell(\hat{\theta}) - 2(\theta - \hat{\theta})' \ell'(\hat{\theta}) - (\theta - \hat{\theta})' \ell''(\hat{\theta})(\theta - \hat{\theta}) \\
&= -2\ell(\hat{\theta}) + (\theta - \hat{\theta})' I(\hat{\theta})(\theta - \hat{\theta}) \; \text{quadratic log-likelihood} \\
L(\theta) &\doteq L(\hat{\theta}) \cdot \exp[-(\theta - \hat{\theta})' I(\hat{\theta})(\theta - \hat{\theta})/2] \; \text{normal likelihood} \\
\pi(\theta|\mathbf{y}) &\doteq c \cdot \exp[-(\theta - \hat{\theta})' I(\hat{\theta})(\theta - \hat{\theta})/2] \; \text{normal posterior}
\end{aligned}
$$

The quadratic form $(\theta - \hat{\theta})' I(\hat{\theta})(\theta - \hat{\theta})$ is (asymptotically) a pivotal (function of data and parameters) which has (asymptotically) a known (χ^2) distribution, Bayesian or frequentist.

Asymptotic distributions

So asymptotically, given the data **y** and a flat prior on θ, we have the posterior distributions:

$$\begin{aligned} \theta &\sim N(\hat{\theta}, I(\hat{\theta})^{-1}), \\ (\theta - \hat{\theta})' I(\hat{\theta})(\theta - \hat{\theta}) &\sim \chi^2_p, \\ D(\theta) &\sim D(\hat{\theta}) + \chi^2_p, \\ L(\theta) &\sim L(\hat{\theta}) \cdot \exp(-\chi^2_p/2). \end{aligned}$$

- The deviance $D(\theta)$ has a shifted χ^2_p distribution, shifted by the frequentist deviance $D(\hat{\theta})$, where p is the dimension of θ.
- The likelihood $L(\theta)$ has a scaled $\exp(-\chi^2_p/2)$ distribution.

With even moderate data sets, likelihoods become extremely small and cause underflow in computations – we evaluate deviances instead.

Posterior deviances

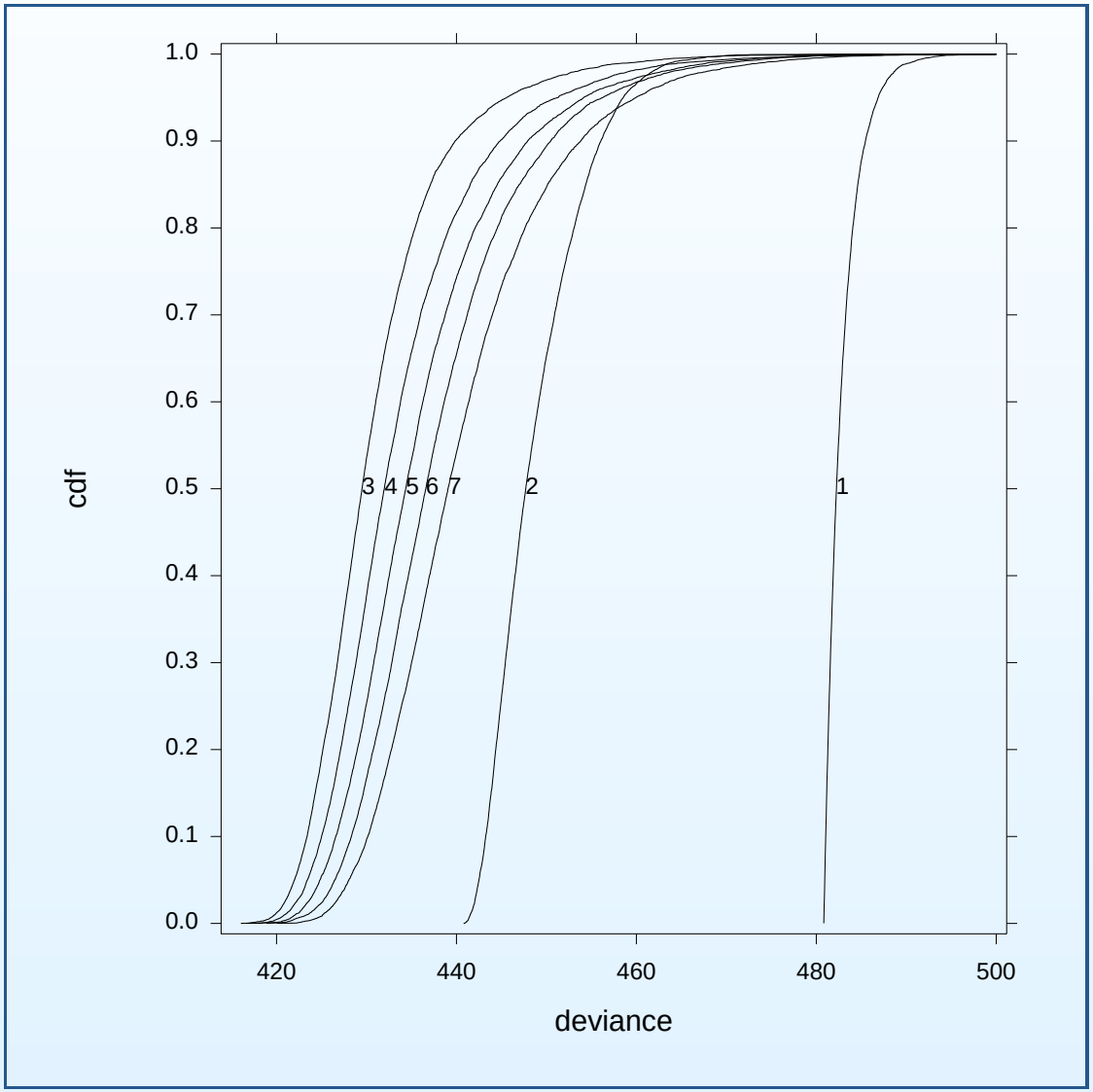
The frequentist deviance is an origin parameter for the posterior deviance distribution: no random draw can give a smaller deviance value than the frequentist deviance.

For the comparison of models with different numbers of components, we compute the sets of posterior deviance draws for $K = 1, 2, ...7$:

$$D_K^{[m]} = -2 \log L_K^{[m]}.$$

The M values for each K define the posterior distributions: we order them to give the empirical cdfs.

Deviance distributions



Interpretation

- The deviance distribution for $K = 1$ is far to the right of the others – the 1-component mixture (single normal distribution) is a very bad fit relative to the others.
- Its deviance distribution is stochastically larger than all the others.
- The fit improves substantially from K = 1 to 2.
- The improvement continues from K = 2 to 3.
- The distribution for $K = 3$ is the stochastically smallest –
- as the number of components increases beyond 3 the deviance distributions move steadily to the right, to larger deviance values (lower likelihoods).
- They also become more diffuse, with decreasing slope.

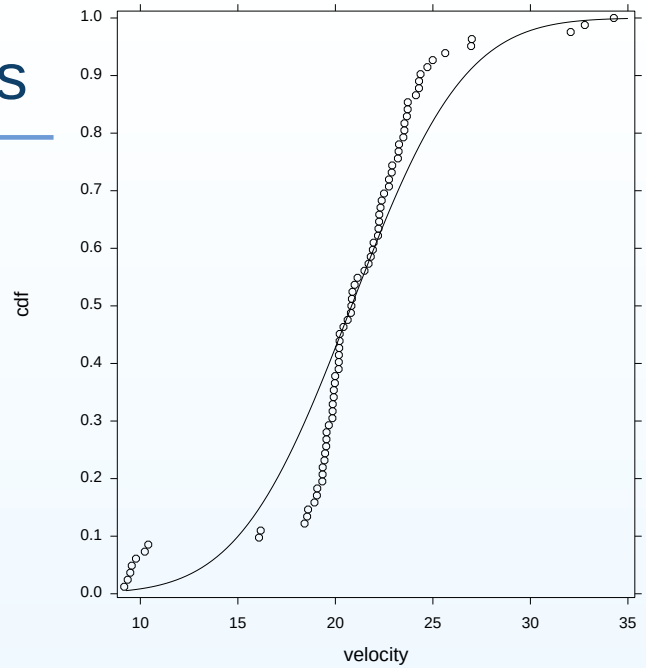
With more parameters there is less information about each parameter, and therefore about the likelihood and deviance, so their posterior distributions become more diffuse.

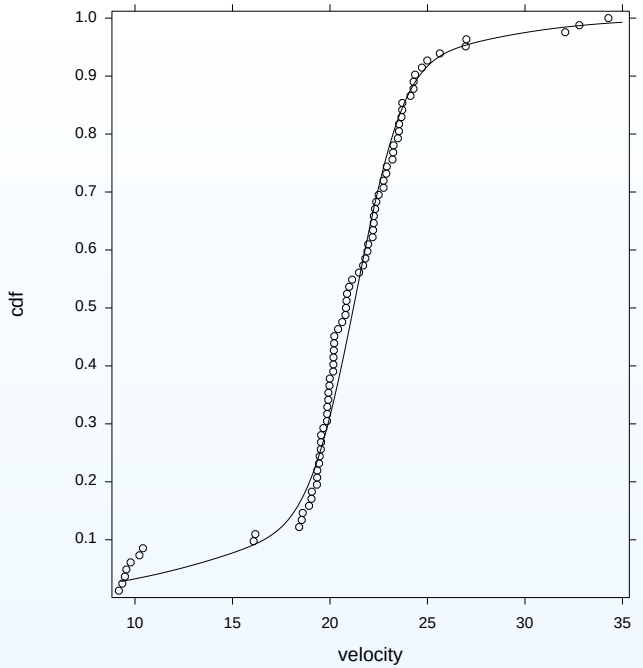
Consistency with data

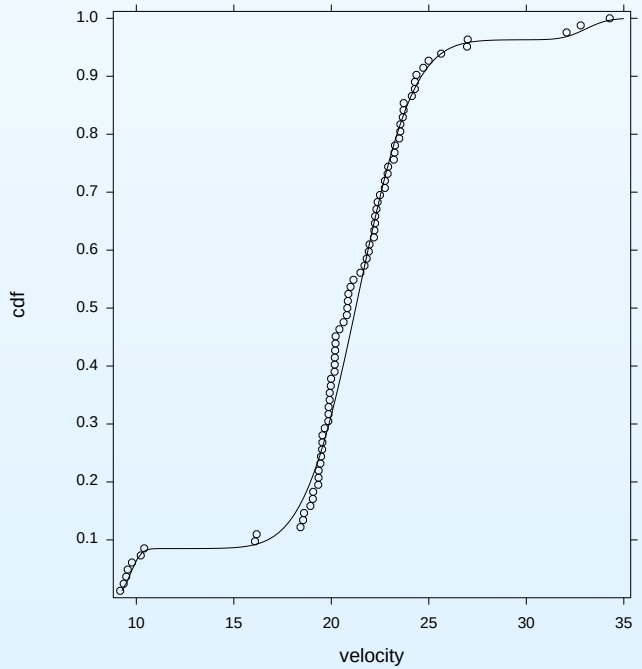
The conclusions from the deviance distributions are consistent with

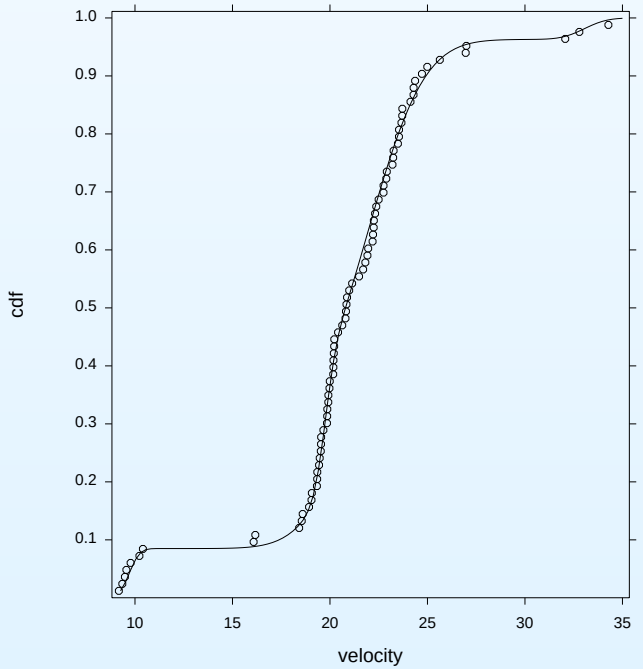
- the appearance of the data empirical cdfs;
- the results of Roeder and Wasserman;
- the DIC results of Celeux et al (whatever the choice of function of the deviance draws, or the penalty, DIC always chose 3 components for the galaxy data).

Graphs


cdf
velocity
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
10
15
20
25
30
35


cdf
velocity
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
10
15
20
25
30
35


cdf
velocity
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
10
15
20
25
30
35


cdf
velocity
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
10
15
20
25
30
35

Does this prove anything?

Bayesian methods for determining the number of mixture components have not previously been supported by simulation studies.

Aitkin, Vu and Francis (2015) report studies based on galaxy-like data sets, and compare model choice by posterior deviances with choice by DIC (which has a penalty on the mean deviance for each model – Spiegelhalter et al 2002).

The deviance distribution comparison was uniformly more accurate than the DIC comparison, especially for large numbers of components.

The 5- and 7-component mixtures were particularly difficult to identify.

Simulations

n	82		164		328		656	
K	DIC	Dev	DIC	Dev	DIC	Dev	DIC	Dev
1	100	100	100	99	100	100	100	100
2	85	98	100	100	100	100	100	97
3	51	99	98	99	100	99	100	99
4	3	9	11	67	30	99	17	99
5	0	18	0	9	0	37	1	89
6	2	9	0	10	56	100	78	100
7	0	1	0	15	4	3	4	32

Percentages of correct model identification in 100 data sets of size n using DIC and posterior deviance

References for this work

Postman, M.J., Huchra, J.P. and Geller, M.J. (1986) Probes of large-scale structures in the Corona Borealis region. The Astronomical Journal 92, 1238–1247.

Roeder, K. (1990) Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. JASA 85, 617–624.

Carlin, B.P. and Chib, S. (1995) Bayesian model choice via Markov Chain Monte Carlo methods. JRSSB 57, 473–484.

Escobar, M.D. and West, M. (1995) Bayesian density estimation and inference using mixtures. JASA 90, 577–588.

Phillips, D.B. and Smith, A.F.M. (1996) Bayesian model comparison via jump diffusions. in Markov chain Monte Carlo in practice eds. Gilks, W.R., S. Richardson and D.J. Spiegelhalter. London: Chapman and Hall.

References for this work

Richardson, S. and Green, P.J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). JRSSB 59, 731–792.

Roeder, K. and Wasserman, L. (1997) Practical Bayesian density estimation using mixtures of normals. JASA 92, 894–902.

Aitkin, M. (2001) Likelihood and Bayesian analysis of mixtures. Statistical Modelling 1, 287–304.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with Discussion). JRSSB 64, 583–639.

Nobile, A. (2004) On the posterior distribution of the number of components in a finite mixture. Annals of Statistics 32, 2044–2073.

References for this work

Celeux, G., Forbes, F., Robert, C. P. and Titterington, D. M. (2006) Deviance information criteria for missing data models. Bayesian Analysis 1, 651–674.

Aitkin, M. (2010) Statistical Inference: an Integrated Bayesian/Likelihood Approach. CRC Press, Boca Raton.

Aitkin, M. (2011) How many components in a finite mixture? pp. 277–292 in Mixtures: Estimation and Applications. eds. K.L. Mengersen, C.P. Robert and D.M. Titterington. Wiley, Chichester.

Gelman, A., Robert, C.P. and Rousseau, J. (2013) Inherent difficulties of non-Bayesian likelihood inference, as revealed by an examination of a recent book by Aitkin. Statistics and Risk Modeling 30, 105–120.

Aitkin, M. (2013) Comments on the review of Statistical Inference. Statistics and Risk Modeling 30, 121–132.

References for this work

Aitkin, M., Vu, D. and Francis, B. (2015) A new Bayesian approach for determining the number of components in a finite mixture. Metron, revision under review.