The University of Newcastle

Kerrie Mengersen

Introduction to

Bayesian Modelling - 5

Armidale 2004

Mixture Models

- Why consider mixture models?
- Trans-dimensional MCMC
- Nonparametric modelling

Bayesian mixture representation

$$y \sim \Sigma_{j=1:k}\, p_j\ \mathrm{f}(\,\mathrm{y}|\theta_j\,)$$

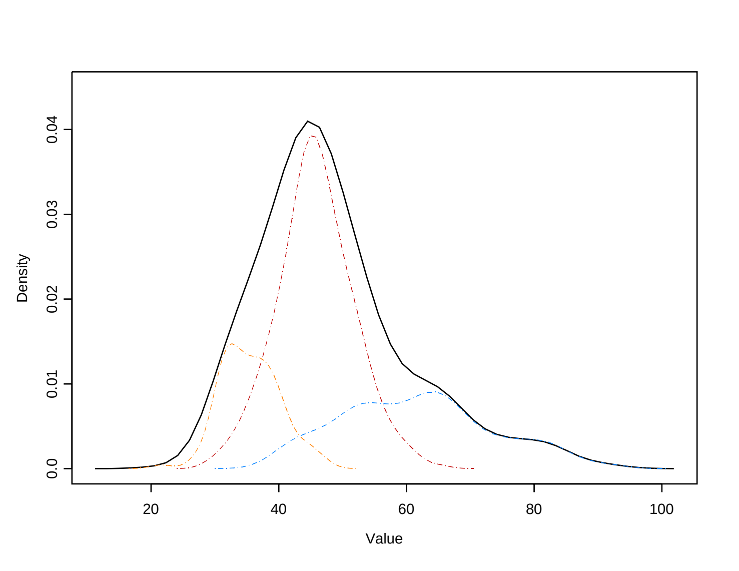
Eg, for mixture of Normals:

$$y \sim \Sigma_{j=1:k}\, p_j\ \mathrm{N}(\,\mu_j\,,\ \tau_j\,)$$

$p \sim$ Dirichlet

$\mu \sim$ Normal

$\tau \sim$ Gamma



Eg, 3 genotypes: qq, qQ, QQ

The p's are the 'weights' assigned to each component. If there are two p's this is like a binomial situation. With more than two components the extension is the multinomial situation. The Dirichlet is a conjugate prior for the multinomial distribution

$p(\theta|\alpha) \propto \Pi\ \theta^{j}\alpha^{j-1}$; setting α=1 for all j gives the Uniform.

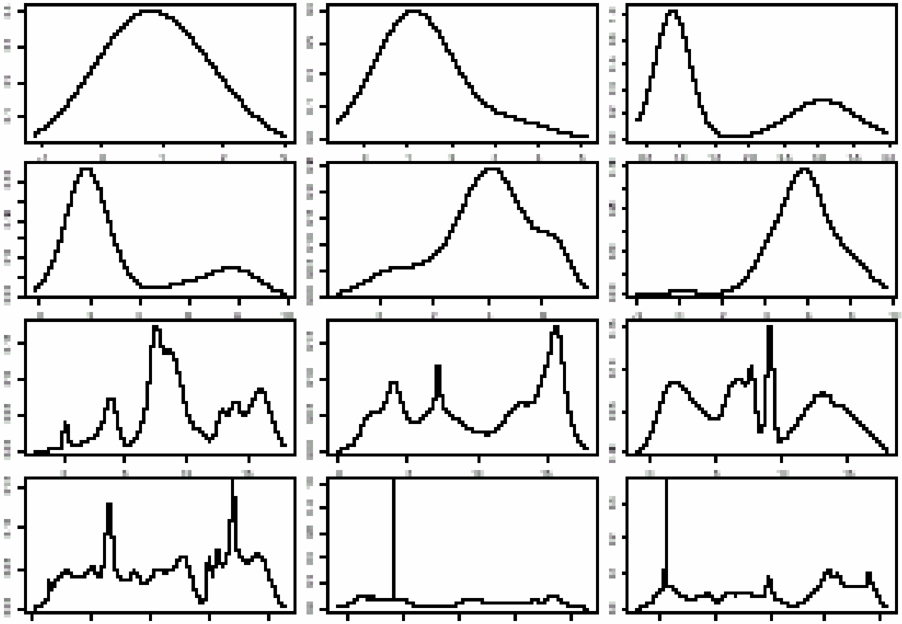
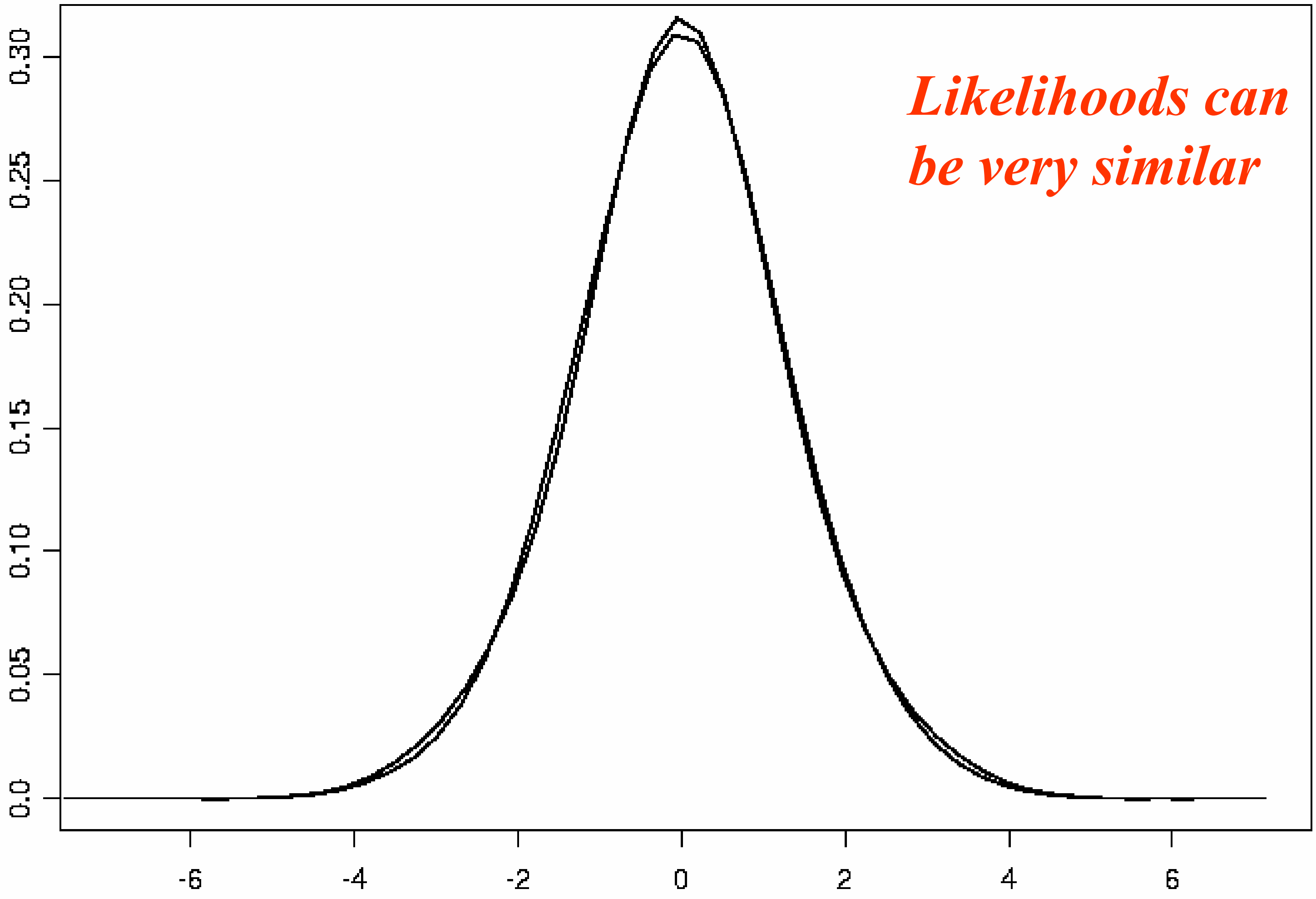


FIGURE 1. Some normal mixture densities for $K = 2$ *(first row)*, $K = 5$ *(second row)*, $K = 25$ *(third row)* and $K = 50$ *(last row)*.

Issues with Mixtures

- Likelihood can be written down but is poorly defined and computationally difficult
- Reparametrisation issues
- Prior modelling is delicate
- Problem of label switching
- Inferences may be problematic
- How to choose the number of components



Likelihoods can be very similar

$0.8N(0,1)+0.1N(-2,1)+0.1N(2,1)$ versus

$0.55N(0,1)+0.225N(-1,1.25)+0.225N(1,1.25)$

The computational problem

- The formulation $y \sim \Sigma_{j=1:k}\, p_j$ f(y|θ_j) means that the likelihood is

$$\mathbb{L}(\underline{\theta}, \underline{p} | \underline{x}) = \prod_{i=1}^{n} \sum_{j=1}^{k} p_j f\left(x_i | \theta_j\right)$$

- This has k^n terms – an unwieldy computation!
- There is a probability $(1\text{-}p_i)^n$ that no observation will be allocated to a component, ie no information from sample to assist estimation, ie likelihood becomes unbounded

Missing data approach

- It is always possible to associate to a r.v. X_i from a mixture of k distributions another r.v. Z_i such that

$$X_i \mid Z_i = z \sim f(x|\theta_z), \qquad Z_i \sim \mathcal{M}_k(1;p_1,\ldots,p_k)$$

where $\mathcal{M}_k(1;p_1,\ldots,p_k)$ is the multinomial distribution.

- Thus Z_i identifies the component of the mixture to which X_i belongs.

Identifiability, or ‘label-switching’

A basic feature of a mixture model is that it is invariant under permutation of the components. Hence the component parameters are not identifiably *marginally*: we cannot distinguish component 1 from component 2 in the likelihood, because they are exchangeable.

This is crucial for both Bayesian inference and computation:

1. Maximisation and exploration of the posterior surface is harder.
2. If the prior on $\theta=(\theta_1,\ldots,\theta_k)$ is exchangeable, the posterior expectation of θ_1 is equal to the posterior expectation of θ_2, etc.
3. Can’t use independent improper priors.

Overcoming label-switching

1. Impose an ‘identifiability constraint’, eg by ordering the means (or the variances or weights): shown to be unstable and undesirable.

Beware!

- This amounts to truncating the original prior distribution. This might radically modify the prior and come close to contradicting it.
- Instead of singling out one mode of the posterior, this might include parts of several modes, so the posterior mean might lie in a very low probability region while the high posterior probability regions are located at the boundaries.
- With many parameters, such ordering is unrealistic.

Overcoming label-switching

2. Introduce a common reference θ_0: scale, location or location-scale parameter. This can now have an improper prior if desired.

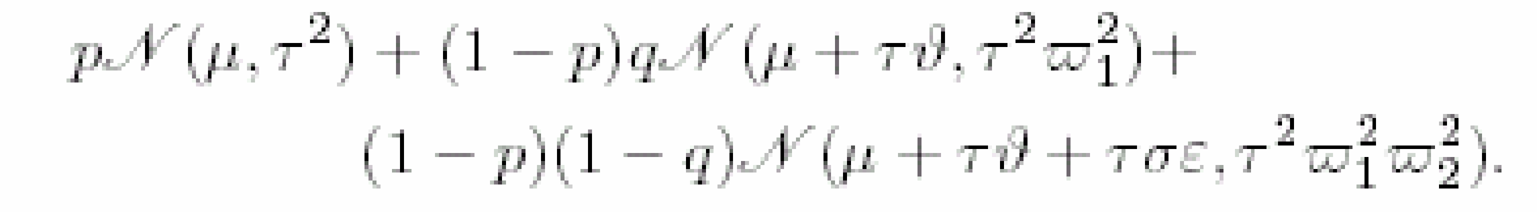
   Define θ_i in terms of *departures* from θ_0.

Eg, Normal case:

- Start from the N(μ,τ^2) distribution.
- Create a two-component mixture

  $$p\ N(\mu,\tau^2) + (1\text{-}p)\ N(\mu+\tau\theta,\tau^2\omega^2)$$

- Three-component mixture:



$$p\mathcal{N}(\mu,\tau^2)+(1-p)q\mathcal{N}(\mu+\tau\vartheta,\tau^2\varpi_1^2)+ \\ (1-p)(1-q)\mathcal{N}(\mu+\tau\vartheta+\tau\sigma\varepsilon,\tau^2\varpi_1^2\varpi_2^2).$$

Gibbs sampling for mixtures

0. *Initialisation:* Choose $\underline{p}^{(0)}$ and $\underline{\theta}^{(0)}$ arbitrarily

For t=1,…

1.1 Allocate observations to components:
Generate $z^{(t)}$ for each observation

1.2 Generate new weights for the components:
Generate $\underline{p}^{(t)}$

1.3 Generate new parameters for each component:
Generate $\underline{\theta}^{(t)}$

Gibbs sampling for mixtures

Consider a 3-component mixture with N observations.

At step *t*:

1.1 To generate z_i:

Use weights p_1,p_2,p_3 and parameters $\theta_1,\theta_2,\theta_3$ from the last iteration:

Calculate $P(z_i^{(t)}=1|\ldots) \propto p_1 \; f(x_i|\theta_1)$

Calculate $P(z_i^{(t)}=2|\ldots) \propto p_2 \; f(x_i|\theta_2)$

Calculate $P(z_i^{(t)}=3|\ldots) \propto p_3 \; f(x_i|\theta_3)$

Gibbs sampling for mixtures

1.2 To generate p:

Use z_i from the last step:

Calculate n_1=no. components allocated to component 1

Calculate n_2=no. components allocated to component 2

Generate p from a Dirichlet distribution

Dirichlet (3; n_1/N, n_2/N)

Gibbs sampling for mixtures

1.3 To generate θ:

For component j, use the observations allocated to that component and estimate parameters using methods discussed previously, ie generate new θ_j from

$p(\theta|\ldots)$

Normal mixture example

$$p\,\mathcal{N}(\mu_1, 1) + (1-p)\,\mathcal{N}(\mu_2, 1)$$ p unknown

Normal prior N(δ,1/λ) on both μ_1 and μ_2

For computation, let s_j^x = sum of the x's allocated to component j

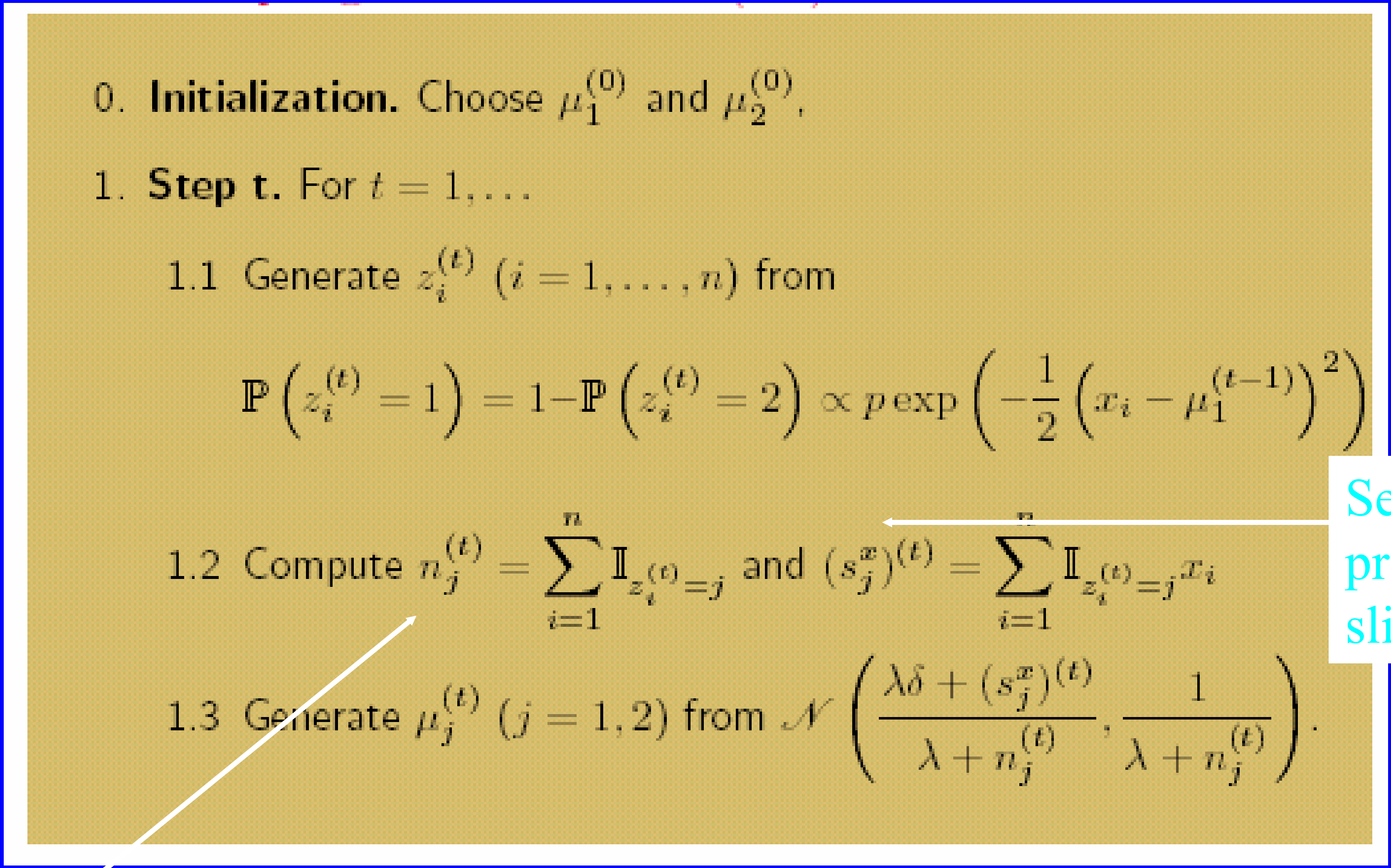
Then μ_1 and μ_2 are independent, given ($\underline{z}$, $\underline{x}$), with conditional distributions

$$\mathcal{N}\left(\frac{\lambda\delta + s_1^x}{\lambda + n_1}, \frac{1}{\lambda + n_1}\right) \quad \text{and} \quad \mathcal{N}\left(\frac{\lambda\delta + s_2^x}{\lambda + n_2}, \frac{1}{\lambda + n_2}\right)$$

Conditional distribution of $\underline{z}$ given (μ_1, μ_2) is

$$\mathbb{P}\left(z_i = 1 | \mu_1, x_i\right) \propto p \exp\left(-0.5\left(x_i - \mu_1\right)^2\right).$$

Gibbs for Normal mixture



0. **Initialization.** Choose $\mu_1^{(0)}$ and $\mu_2^{(0)}$.

1. **Step t.** For $t = 1, \ldots$

   1.1 Generate $z_i^{(t)}$ $(i = 1, \ldots, n)$ from

$$\mathbb{P}\left(z_i^{(t)} = 1\right) = 1 - \mathbb{P}\left(z_i^{(t)} = 2\right) \propto p \exp\left(-\frac{1}{2}\left(x_i - \mu_1^{(t-1)}\right)^2\right)$$

   1.2 Compute $n_j^{(t)} = \sum_{i=1}^{n} \mathbb{I}_{z_i^{(t)}=j}$ and $(s_j^x)^{(t)} = \sum_{i=1}^{n} \mathbb{I}_{z_i^{(t)}=j} x_i$

   1.3 Generate $\mu_j^{(t)}$ $(j = 1, 2)$ from $\mathcal{N}\left(\frac{\lambda\delta + (s_j^x)^{(t)}}{\lambda + n_j^{(t)}}, \frac{1}{\lambda + n_j^{(t)}}\right)$.

See previous slide

$n_j^{(t)}$ is number of observations allocated to component j

Example

.7N(0,1) + .3N(2.5,1)

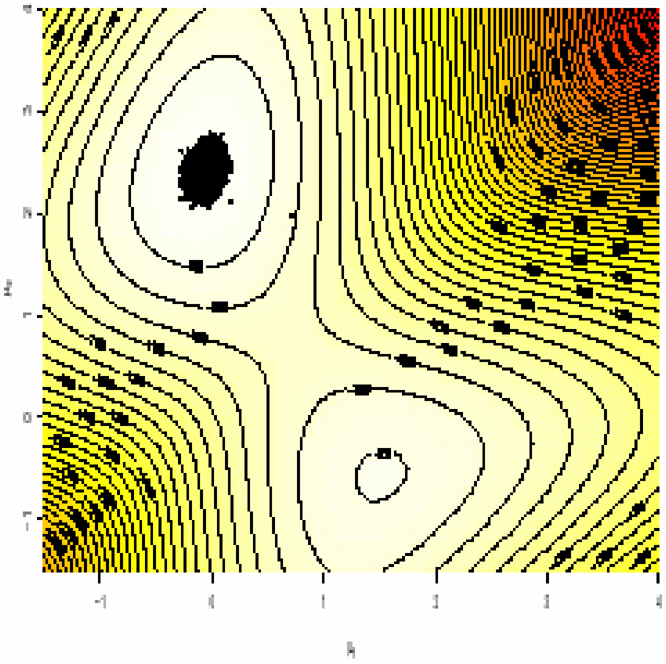


FIGURE 12. Log-posterior surface and the corresponding Gibbs sample for the model (1.7), based on 10,000 iterations.

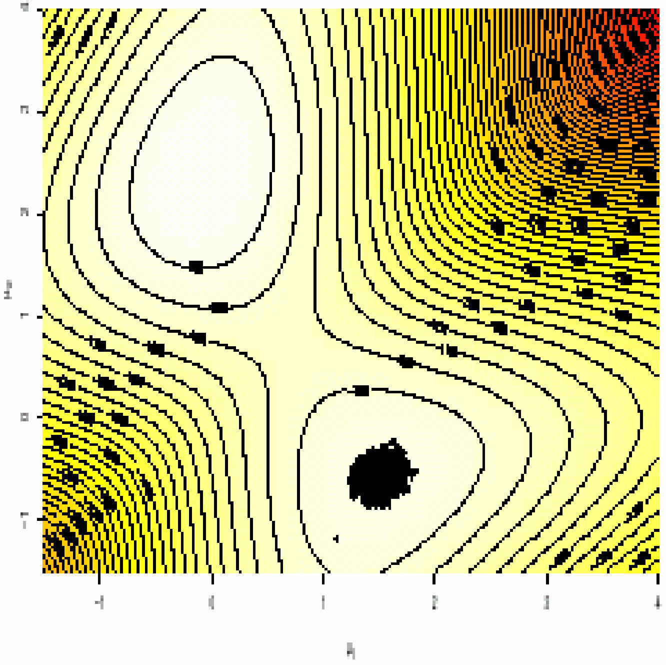


FIGURE 13. Same graph, when initialised close to the second and lower mode, based on 10,000 iterations.

Dependent on initial conditions: can ‘get stuck’ in small mode!

M-H escapes trapping states better

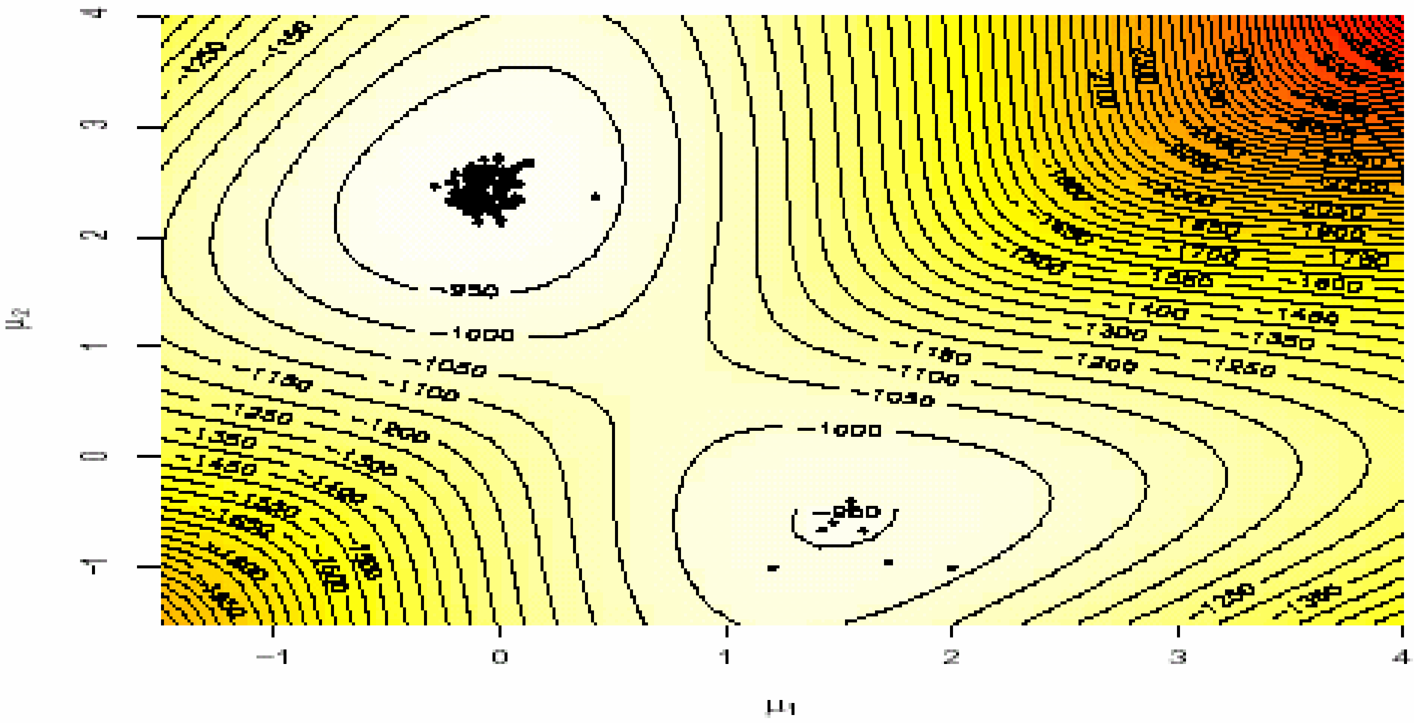


FIGURE 17. Track of a 10, 000 iterations random walk Metropolis–Hastings sample on the posterior surface, the starting point is equal to (2,-1). The scale of the random walk ζ^2 is equal to 1.

Effect of ‘tuning’

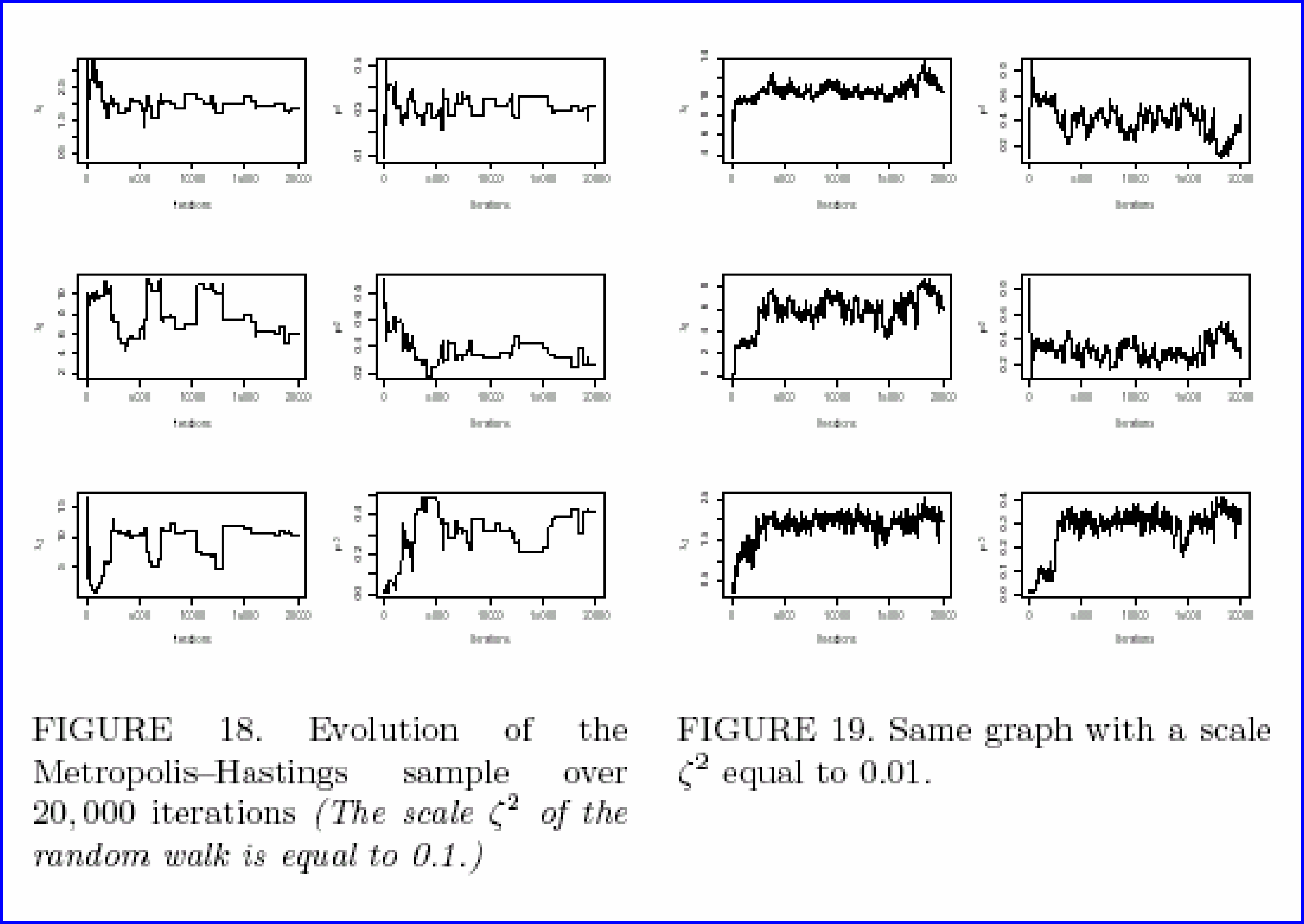
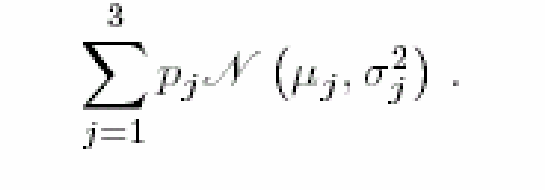


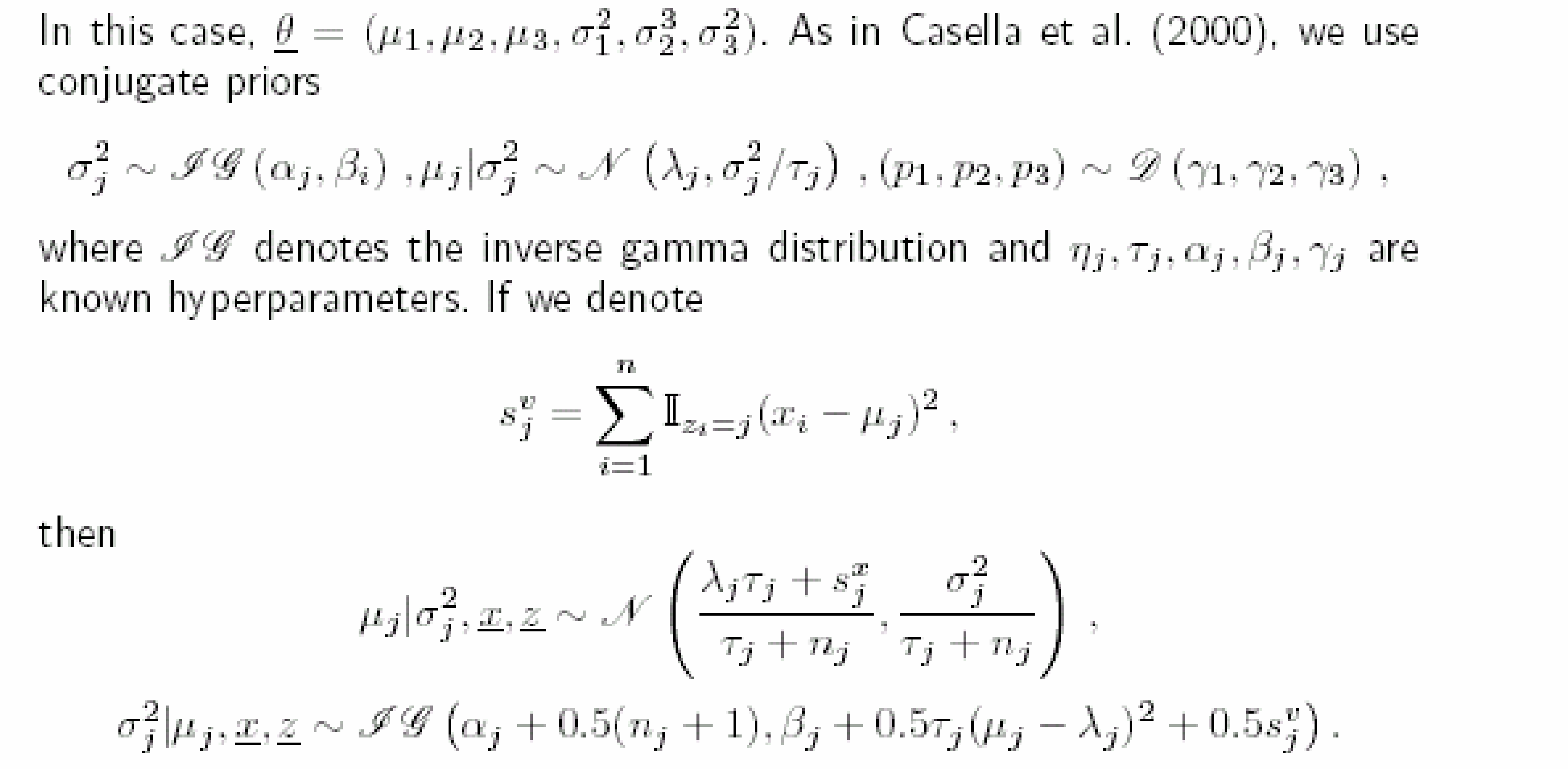
FIGURE 18. Evolution of the Metropolis–Hastings sample over 20,000 iterations *(The scale ζ^2 of the random walk is equal to 0.1.)*

FIGURE 19. Same graph with a scale ζ^2 equal to 0.01.

3-component Normal mixture



$$\sum_{j=1}^{3} p_j \mathscr{N}\left(\mu_j, \sigma_j^2\right) .$$



In this case, $\underline{\theta} = (\mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^3, \sigma_3^2)$. As in Casella et al. (2000), we use conjugate priors

$$\sigma_j^2 \sim \mathscr{IG}\left(\alpha_j, \beta_i\right), \mu_j|\sigma_j^2 \sim \mathscr{N}\left(\lambda_j, \sigma_j^2/\tau_j\right), (p_1, p_2, p_3) \sim \mathscr{D}\left(\gamma_1, \gamma_2, \gamma_3\right),$$

where $\mathscr{IG}$ denotes the inverse gamma distribution and $\eta_j, \tau_j, \alpha_j, \beta_j, \gamma_j$ are known hyperparameters. If we denote

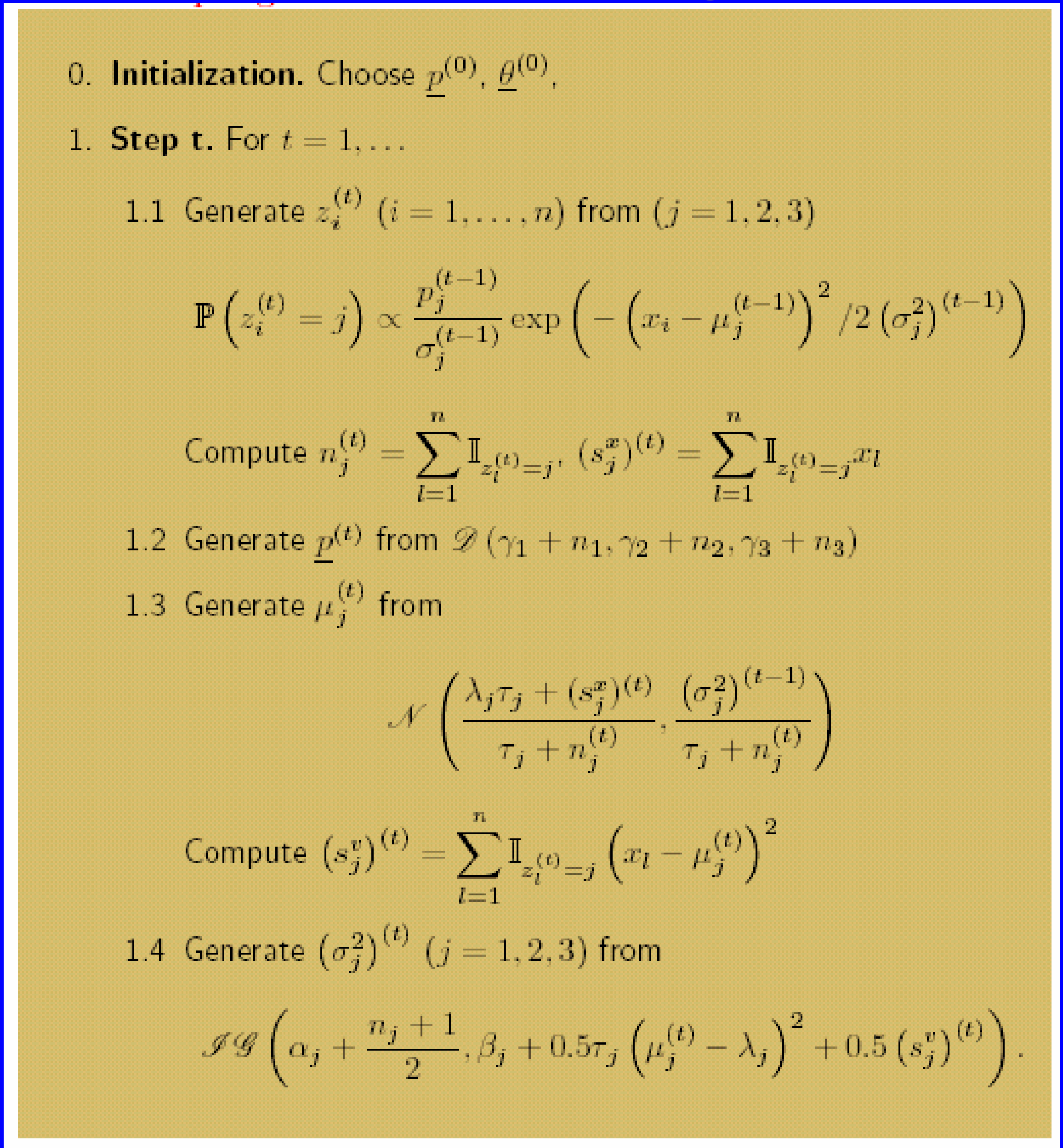
$$s_j^v = \sum_{i=1}^{n} \mathbb{I}_{z_i=j}(x_i - \mu_j)^2 ,$$

then

$$\mu_j|\sigma_j^2, \underline{x}, \underline{z} \sim \mathscr{N}\left(\frac{\lambda_j \tau_j + s_j^x}{\tau_j + n_j}, \frac{\sigma_j^2}{\tau_j + n_j}\right),$$

$$\sigma_j^2|\mu_j, \underline{x}, \underline{z} \sim \mathscr{IG}\left(\alpha_j + 0.5(n_j + 1), \beta_j + 0.5\tau_j(\mu_j - \lambda_j)^2 + 0.5 s_j^v\right).$$

MCMC algorithm



0. **Initialization.** Choose $\underline{p}^{(0)}$, $\underline{\theta}^{(0)}$,

1. **Step t.** For $t = 1, \ldots$

   1.1 Generate $z_i^{(t)}$ $(i = 1, \ldots, n)$ from $(j = 1, 2, 3)$

$$\mathbb{P}\left(z_i^{(t)} = j\right) \propto \frac{p_j^{(t-1)}}{\sigma_j^{(t-1)}} \exp\left(-\left(x_i - \mu_j^{(t-1)}\right)^2 / 2\left(\sigma_j^2\right)^{(t-1)}\right)$$

   Compute $n_j^{(t)} = \sum_{l=1}^{n} \mathbb{I}_{z_l^{(t)}=j}$, $(s_j^x)^{(t)} = \sum_{l=1}^{n} \mathbb{I}_{z_l^{(t)}=j} x_l$

   1.2 Generate $\underline{p}^{(t)}$ from $\mathscr{D}\left(\gamma_1 + n_1, \gamma_2 + n_2, \gamma_3 + n_3\right)$

   1.3 Generate $\mu_j^{(t)}$ from

$$\mathscr{N}\left(\frac{\lambda_j \tau_j + (s_j^x)^{(t)}}{\tau_j + n_j^{(t)}}, \frac{(\sigma_j^2)^{(t-1)}}{\tau_j + n_j^{(t)}}\right)$$

   Compute $(s_j^v)^{(t)} = \sum_{l=1}^{n} \mathbb{I}_{z_l^{(t)}=j} \left(x_l - \mu_j^{(t)}\right)^2$

   1.4 Generate $(\sigma_j^2)^{(t)}$ $(j = 1, 2, 3)$ from

$$\mathscr{IG}\left(\alpha_j + \frac{n_j + 1}{2}, \beta_j + 0.5\tau_j \left(\mu_j^{(t)} - \lambda_j\right)^2 + 0.5 \left(s_j^v\right)^{(t)}\right).$$

After $20,000$ iterations, the Gibbs sample is quite stable (although more detailed convergence assessment is necessary and the algorithm fails to visit the permutation modes) and, using the $5,000$ last reordered iterations, we find that the posterior mean estimations of μ_1, μ_2, μ_3 are equal to $9.5, 21.4, 26.8$, those of $\sigma_1^2, \sigma_2^2, \sigma_3^2$ are equal to $1.9, 6.1, 34.1$ and those of p_1, p_2, p_3 are equal to $0.09, 0.85, 0.06$. Figure 15 shows the histogram of the data along with the estimated (plug-in) density.

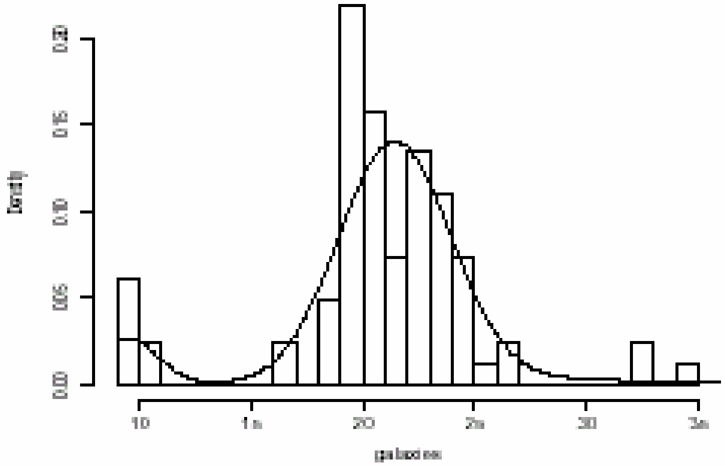


FIGURE 15. Histogram of the velocity of 82 galaxies against the plug-in estimated 3 component mixture, using a Gibbs sampler.

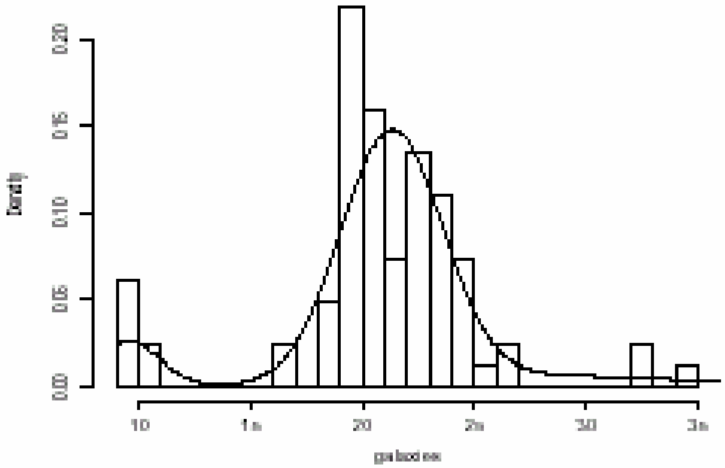


FIGURE 16. Same graph, when using a Metropolis–Hastings algorithm with $\zeta^2 = .01$.

BUGS code

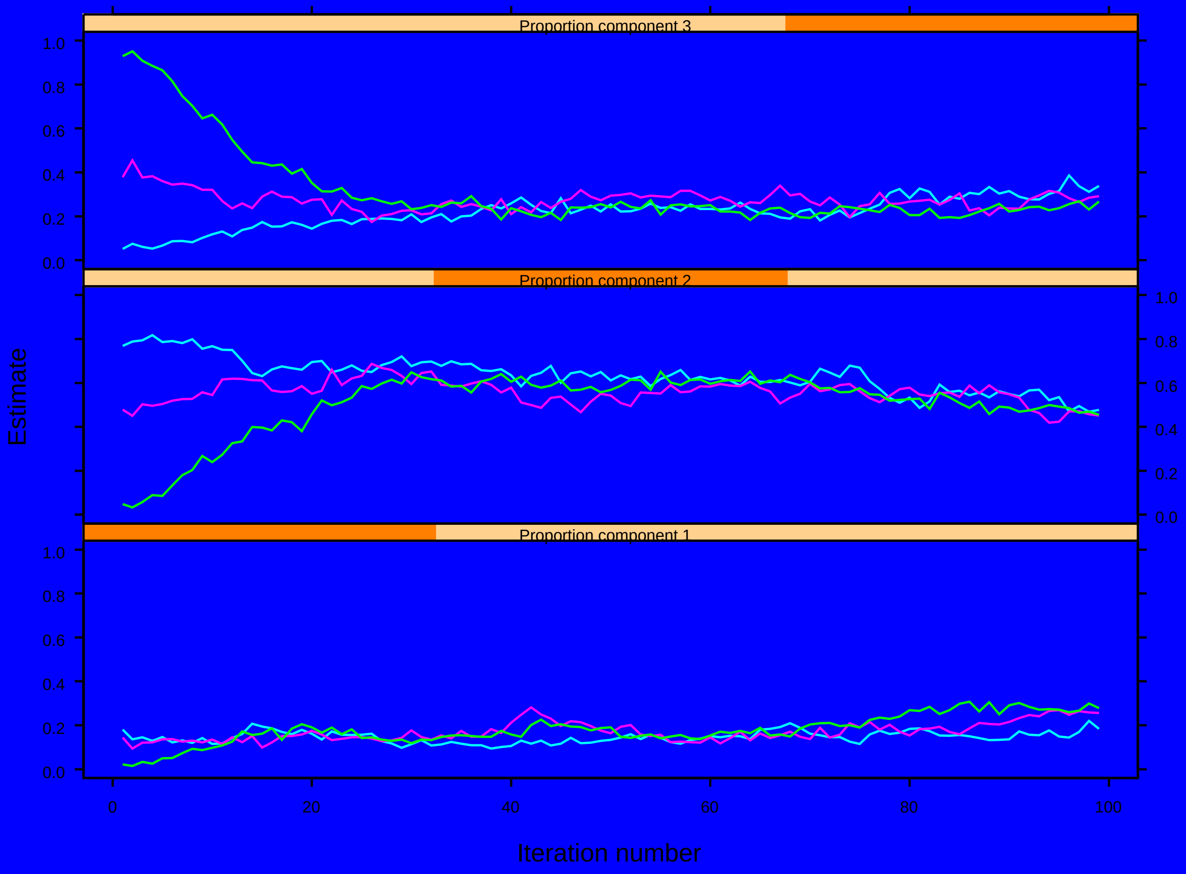
```
model
    {
            for( i in 1 : N ) {
                        y[i] ~ dnorm(mu[i], tau[T[i]])
                        mu[i] <- lambda[T[i]]
                        T[i] ~ dcat(P[])
            }
            P[1:3] ~ ddirch(alpha[])
            lambda[3] ~ dnorm(0.0, 1.0E-6)I(lambda[2], )
            lambda[2] ~ dnorm(0.0, 1.0E-6)I(lambda[1], )
            lambda[1] ~ dnorm(0.0, 1.0E-6)
            tau[3] ~ dgamma(0.001, 0.001)  sigma[3] <- 1 / sqrt(tau[3])
            tau[2] ~ dgamma(0.001, 0.001) sigma[2] <- 1 / sqrt(tau[2])
            tau[1] ~ dgamma(0.001, 0.001) sigma[1] <- 1 / sqrt(tau[1])
}
```

BUGS Results

A 1000 update burn in followed by a further 20000 updates using 3 chains gave the parameter estimates

Param	Mean	SD	Actual
λ_1	0.18	0.053	0.18
λ_2	0.57	0.068	0.55
λ_3	0.25	0.042	0.27
μ_1	34.8	1.333	35.6
μ_2	45.9	0.699	45.8
μ_3	63.3	2.263	61.7
σ_1	4.03	0.683	4.59
σ_2	5.32	0.535	5.93
σ_3	11.34	1.106	11.89

Trace Plots for some parameters



Bayesians in the Night (Strangers in the Night)

Bayesians in the night
with exchangeable glances
Assessing in the night
the prior chances
We'd be sharing risks
before the night is through.

Something in your prior
was so exciting
Something in your data
was so inviting
Something in your posterior
told me I must have you.