Density estimation with an anticipated number of modes

Jochen Einbeck

Department of Mathematical Sciences

Durham University

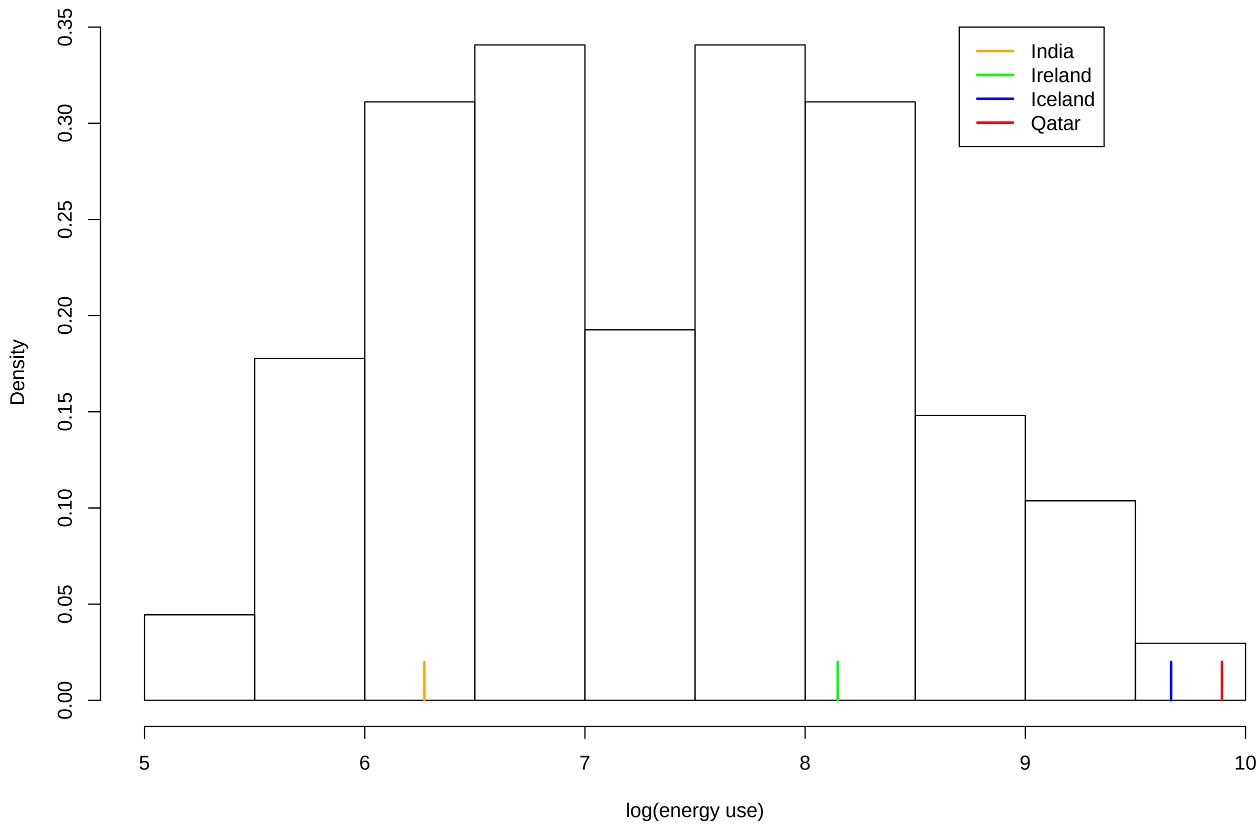
`jochen.einbeck@durham.ac.uk`

joint work with James Taylor (Durham University)

Galway, 19th of May 2011

Motivation: Energy data

- Energy consumption of $n = 135$ countries, in kg of oil equivalent per capita, in the year 2007.
- Plotted is histogram of log- energy consumption, with four exemplary countries highlighted.

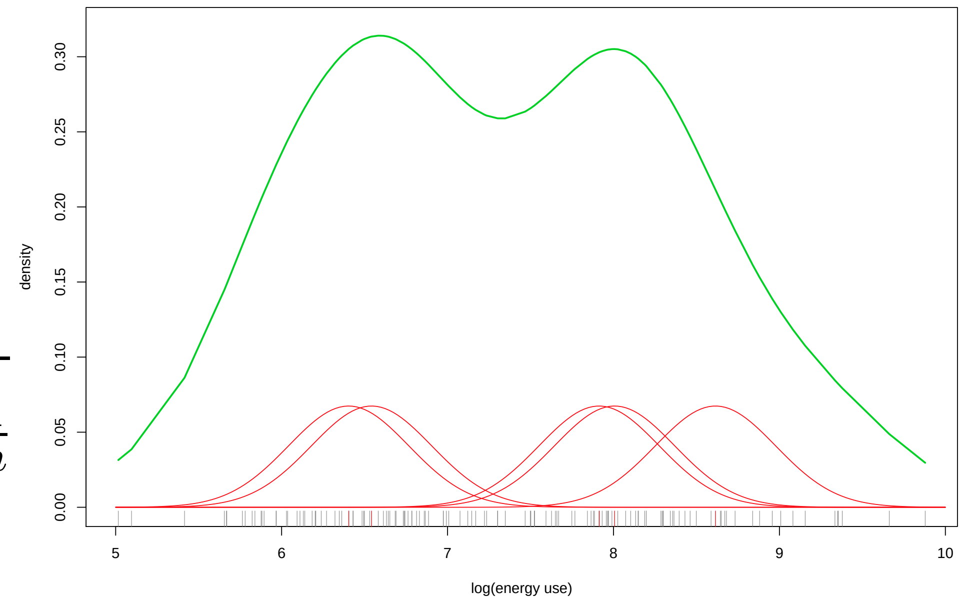


Kernel density estimation

- Alternative to Histogram: Density Estimation
- The kernel density estimator

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)$$

estimates the density by re-distributing the point mass $\frac{1}{n}$ smoothly to its vicinity.



- Popular choice of K: Gaussian density.

Bandwidth selection

- Choose h by minimizing the asymptotic integrated MSE,

$$\begin{aligned}\int \mathsf{MSE}(x)\,dx &= \int \left[\mathsf{Bias}^2(\hat{f}(x)) + \mathsf{Var}(\hat{f}(x))\right] dx = \\ &\approx \frac{\kappa_1 h^4}{4}\int (f''(x))^2\,dx + \frac{\kappa_2}{nh}\end{aligned}$$

yielding

$$h_{opt} = \kappa_0 \left[\int (f''(x))^2\,dx\right]^{-1/5} n^{-1/5}$$

(where $\kappa_j, j = 0, 1, 2$ are constants only depending on K).

Bandwidth selection

- Choose h by minimizing the asymptotic integrated MSE,

$$\begin{aligned}\int \mathsf{MSE}(x)\,dx &= \int \left[\mathsf{Bias}^2(\hat{f}(x)) + \mathsf{Var}(\hat{f}(x))\right] dx = \\ &\approx \frac{\kappa_1 h^4}{4}\int (f''(x))^2\,dx + \frac{\kappa_2}{nh}\end{aligned}$$

  yielding

$$h_{opt} = \kappa_0 \left[\int (f''(x))^2\,dx\right]^{-1/5} n^{-1/5}$$

  (where $\kappa_j, j = 0, 1, 2$ are constants only depending on K).

- Problem: $\int (f''(x))^2\,dx$ unknown !

Normal reference bandwidth selection

- Idea (Silverman, 1986): Replace $\int (f''(x))^2\,dx$ by that value that would be obtained for a normal density $\phi_{0,\sigma} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/(2\sigma^2)}$ with the same variance as f ("normal reference").
- One finds

$$\int (\phi''_{0,\sigma}(x))^2\,dx = \frac{1}{\sigma^5}\int (\phi''_{0,1}(x))^2\,dx = \frac{3}{8\sqrt{\pi}}\sigma^{-5}.$$

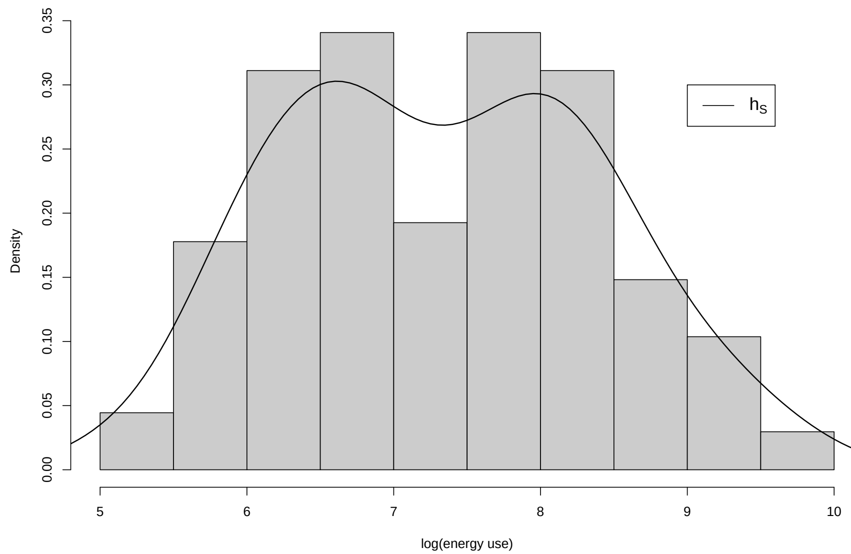
Using $\kappa_0 = 0.776$ for a Gaussian kernel K, one gets

$$h_S = 1.06\sigma n^{-1/5},$$

where σ is estimated using the sample standard deviation, s.

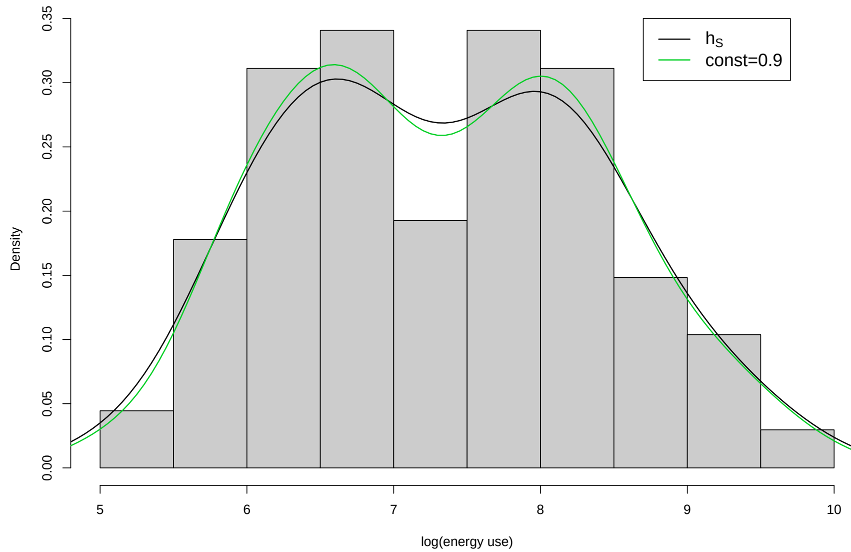
Normal reference bandwidth selection (cont.)

- For the energy data, $s = 1.074$, $n = 135$, so $h = 1.06 \times 1.074 \times 135^{-1/5} = 0.43$.
- Resulting fit looks not too bad, but method tends to oversmooth if the data are multimodal.



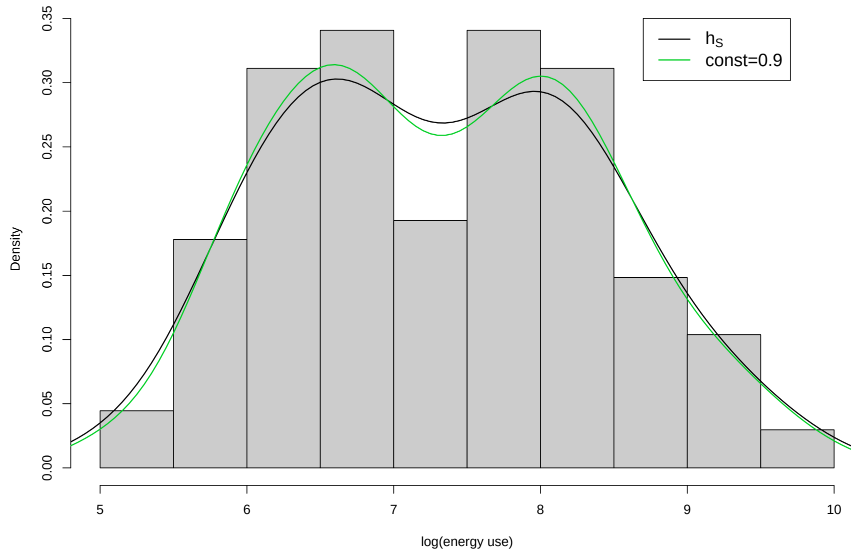
Normal reference bandwidth selection (cont.)

- For the energy data, $s = 1.074$, $n = 135$, so $h = 1.06 \times 1.074 \times 135^{-1/5} = 0.43$.
- Resulting fit looks not too bad, but method tends to oversmooth if the data are multimodal.
- Ad-hoc fix by Silverman: Replace the constant 1.06 with the smaller value 0.9.



Normal reference bandwidth selection (cont.)

- For the energy data, $s = 1.074$, $n = 135$, so $h = 1.06 \times 1.074 \times 135^{-1/5} = 0.43$.
- Resulting fit looks not too bad, but method tends to oversmooth if the data are multimodal.
- Ad-hoc fix by Silverman: Replace the constant 1.06 with the smaller value 0.9.
- Sought:



A systematic rule or justification
how to reduce the constant 1.06 under multimodality.

Reference to a Gaussian mixture

- Obviously, the issue is with $D_f \equiv \int (f''(x))^2 \, dx$.
- If the data are multimodal, then reference to a normal distribution will give a wrong result.
- Mathematical exercise: What happens if we refer to a mixture of normals instead?
  - Postulating say, m, modes, this gives the density

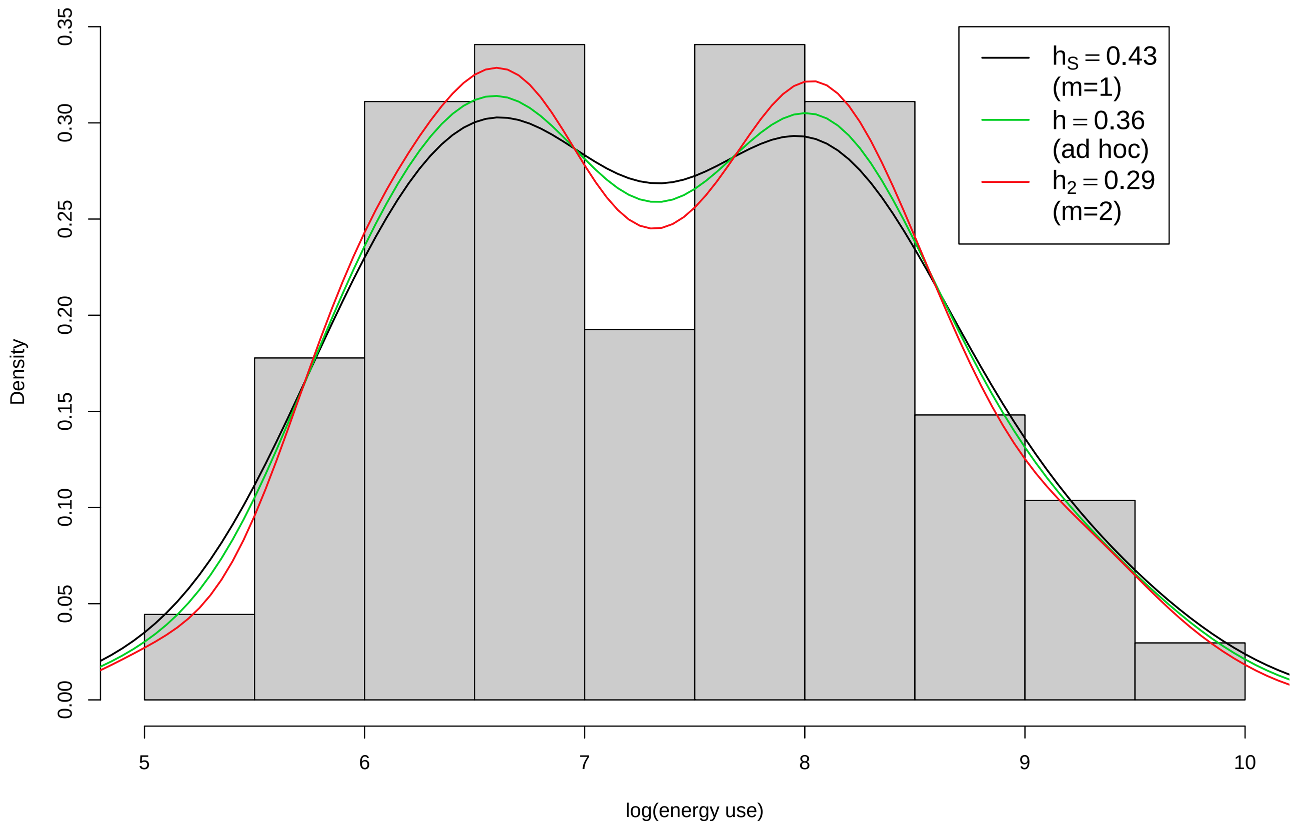
$$\varphi_m(x) = p_1 \phi_{\mu_1,\sigma_1}(x) + \ldots + p_m \phi_{\mu_m,\sigma_m}(x)$$

  - The parameters p_j, μ_j, σ_j can be estimated through the EM algorithm (for instance, R package **npmlreg**).
  - The integral $D_{\varphi_m} = \int (\varphi''_m(x))^2(x) \, dx$ can then be solved numerically (for instance, using Mathematica).
  - Finally,

$$h_m = \kappa_0 D_{\varphi_m}^{-1/5} n^{-1/5}.$$

Reference to a Gaussian mixture (cont.)

- For the energy data with $m = 2$, one obtains $D_{\varphi_2} = 0.96$, so $h_2 = 0.29$.
  - For comparison, for $m = 1$, $D_{\varphi_1} = 0.15$.
- Resulting density estimate:



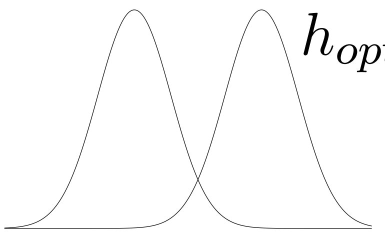
Shortcut

- This seems rather useless: Nobody will take the trouble of fitting a mixture just in order to produce a bandwidth for a kernel density estimate (especially, as the mixture produces a density estimate itself!).
- However, we can simplify things considerably.
  - Assume an equal mixture of m components of equal s.dev. σ, which are all separated by a distance d.
  - Then tedious calculation yields

$$h_{opt} \approx 1.06 m^{-4/5} s \frac{2\sqrt{3}}{d\sqrt{1 + (\frac{12}{d^2} - 1)/m^2}} n^{-1/5}$$

Shortcut

- This seems rather useless: Nobody will take the trouble of fitting a mixture just in order to produce a bandwidth for a kernel density estimate (especially, as the mixture produces a density estimate itself!).
- However, we can simplify things considerably.
  - Assume an equal mixture of m components of equal s.dev. σ, which are all separated by a distance d.
  - Then tedious calculation yields



$$h_{opt} \approx 1.06 m^{-4/5} s \frac{2\sqrt{3}}{d\sqrt{1 + (\frac{12}{d^2} - 1)/m^2}} n^{-1/5}$$

  - For $d = 2\sqrt{3}$, corresponding to well-separated modes, this boils down to

$$h_m = 1.06 m^{-4/5} s n^{-1/5}$$

Shortcut (cont.)

- Rule of thumb:

  For $m-$modal distributions,
  multiply the normal–reference–bandwidth with $m^{-4/5}$.

- Specifically, anticipating m modes, the "mixture-of–normals" reference bandwidths are given by

$$h_m = c(m)sn^{-1/5}$$

with

m	1	2	3	4	5	6	7
$c(m)$	1.06	0.61	0.44	0.35	0.29	0.25	0.22

Shortcut (cont.)

- Rule of thumb:

  For $m-$modal distributions,
  multiply the normal–reference–bandwidth with $m^{-4/5}$.

- Specifically, anticipating m modes, the “mixture-of–normals” reference bandwidths are given by

$$h_m = c(m)sn^{-1/5}$$

  with

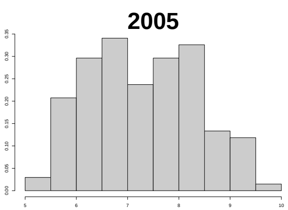
m	1	2	3	4	5	6	7
$c(m)$	1.06	0.61	0.44	0.35	0.29	0.25	0.22

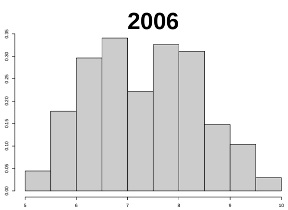
- Note: Except for $m = 1$, all values « 0.9 !!

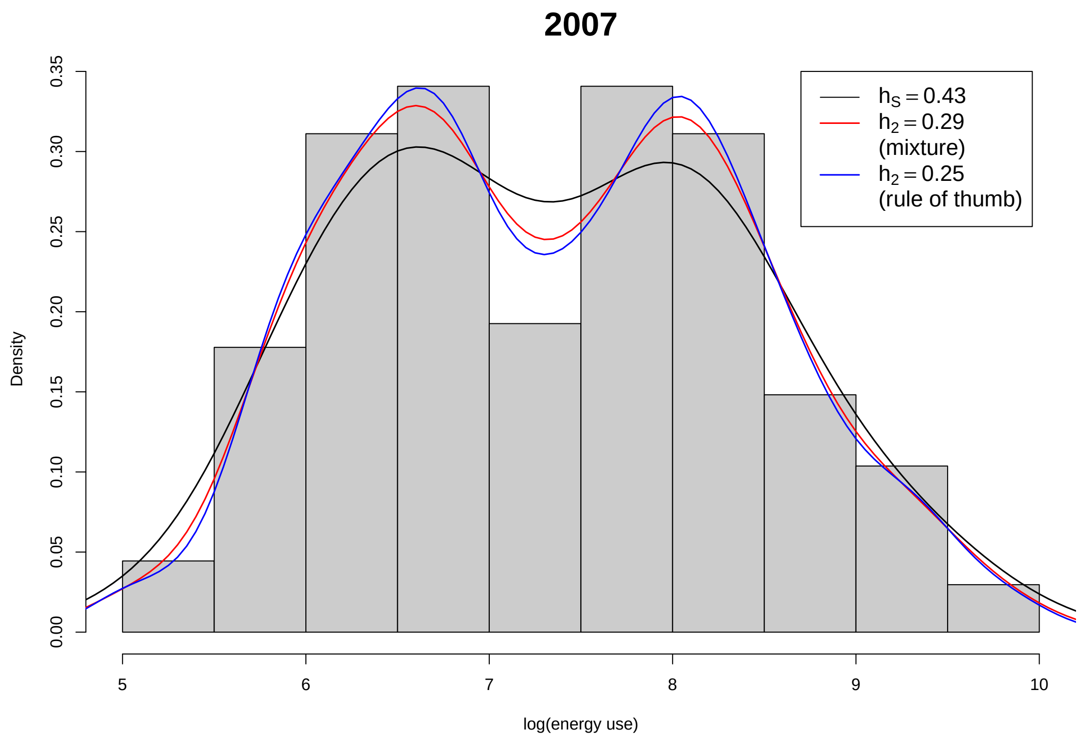
Back to energy data

- Anticipating $m = 2$ modes, for instance from background or expert knowledge, such as the shape of the distribution from previous years, the rule of thumb-bandwidth selector gives

$$h_2 = 1.06 \times 2^{-4/5} \times 1.074 \times 135^{-4/5} = 0.25.$$



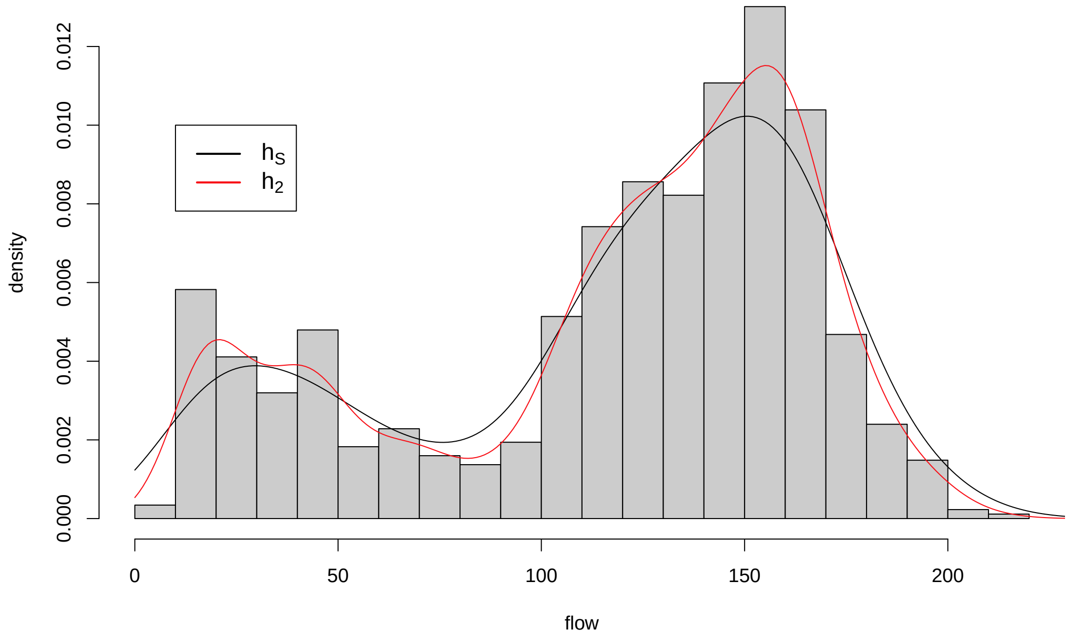




Traffic data

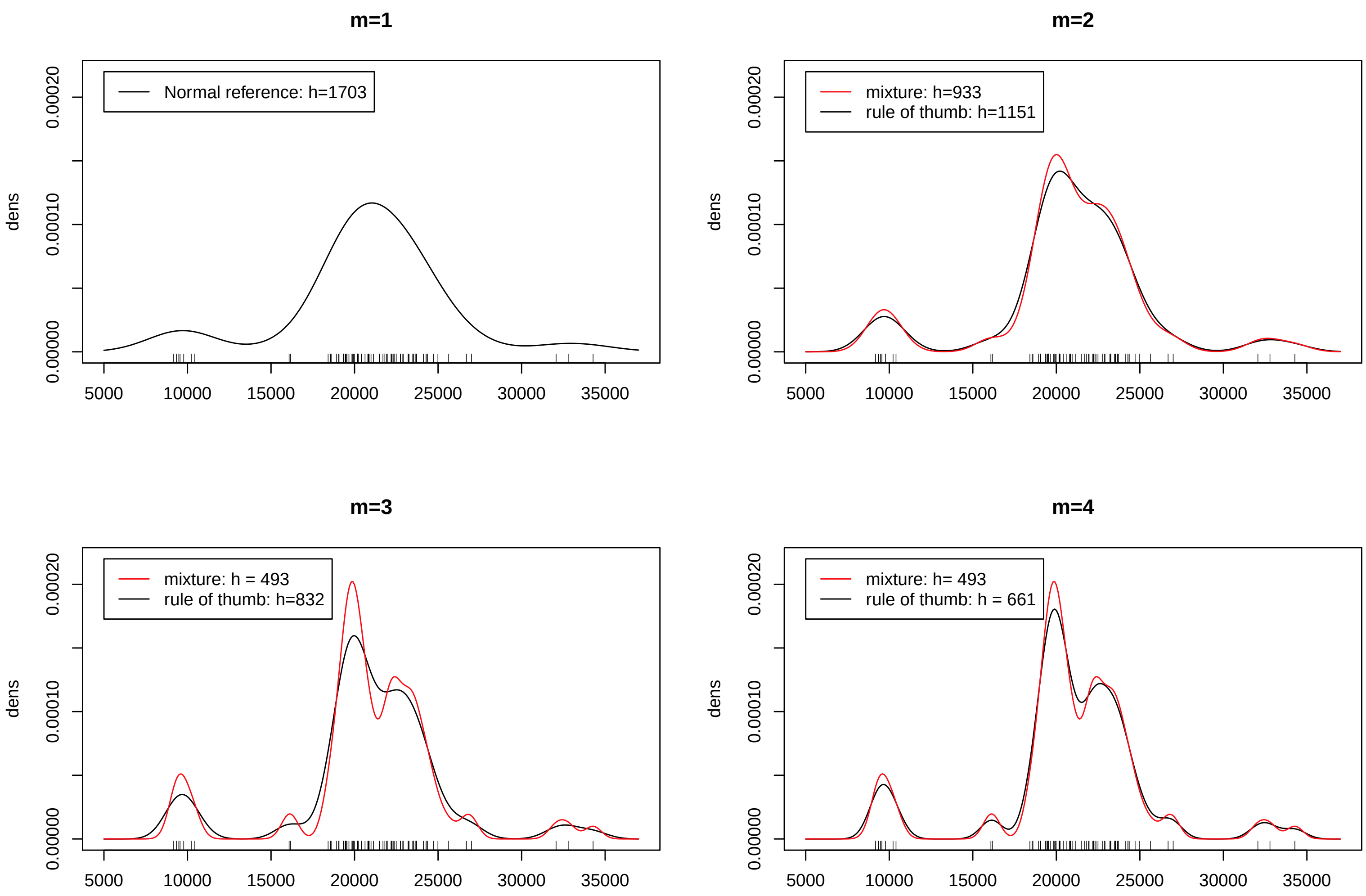
- $n = 876$ measurements of traffic flow (veh/5min) 10-12/07/07 on Californian freeway.
- Normal reference gives $h_S = 13.90$.
- Indeed, traffic engineers might expect at least two modes (freeflow, busy traffic).

- So, $m = 2$ gives $h_2 = 2^{-4/5} \times h_S = 7.98$.
- Anticipating $m = 2$ unveils a third mode!



Galaxy data

- Velocities in km/sec of $n = 82$ galaxies.



Conclusion

- For situations where background/expert knowledge on the modality is available, this information can be used to find a bandwidth of corresponding resolution.
- Rather than needing to estimate D_f accurately through a fitted mixture, a simple rule of thumb criterion can be applied.
- There is no guarantee that the number of modes obtained using this bandwidth corresponds *exactly* to the number of anticipated modes — in fact, it will often be larger.
- General message to take away: With an increasing number of modes, the bandwidth should be reduced by the magnitude $m^{-4/5}$.

Conclusion

- For situations where background/expert knowledge on the modality is available, this information can be used to find a bandwidth of corresponding resolution.
- Rather than needing to estimate D_f accurately through a fitted mixture, a simple rule of thumb criterion can be applied.
- There is no guarantee that the number of modes obtained using this bandwidth corresponds *exactly* to the number of anticipated modes — in fact, it will often be larger.
- General message to take away: With an increasing number of modes, the bandwidth should be reduced by the magnitude $m^{-4/5}$.

References

Silverman (1986): *Density Estimation*. Chapman & Hall/CRC.