Making the EM algorithm for NPML estimation less sensitive to tuning parameters

Jochen Einbeck and John Hinde

National University of Ireland, Galway

CASI, 20th May 2005

Work supported by science foundation ireland
fondúireacht eolaíochta éireann

NPML estimation

Generalized linear model with random effect:

$$\mu_i \equiv E(y_i|z_i, \beta) = h(\eta_i) \equiv h(x_i'\beta + z_i).$$

The marginal likelihood can be approximated by a finite mixture (Laird, 1978)

$$L(\beta, g(z)) = \prod_{i=1}^{n} \int f(y_i|z_i, \beta) g(z_i)\, dz_i \approx \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} f(y_i|z_k, \beta)\pi_k \right\}$$

with mass points z_k and masses π_k, where no parametric assumption about the random effect distribution $g(z_i)$ is made.

$\Longrightarrow$ Nonparametric maximum likelihood (NPML).

A special case: Fitting finite Gaussian mixtures

Assume data $Y_1, \ldots, Y_n$ sampled from a population Y, which is presumed to have the structure of a finite Gaussian mixture, i.e.

$$f(y|(z_k, \pi_k, \sigma_k)_{k=1,\ldots,K}) = \sum_{k=1}^{K} \pi_k f(y|z_k, \sigma_k^2)$$

where $f(y|z_k, \sigma_k^2)$ is a normal density with mean z_k and standard deviation σ_k.

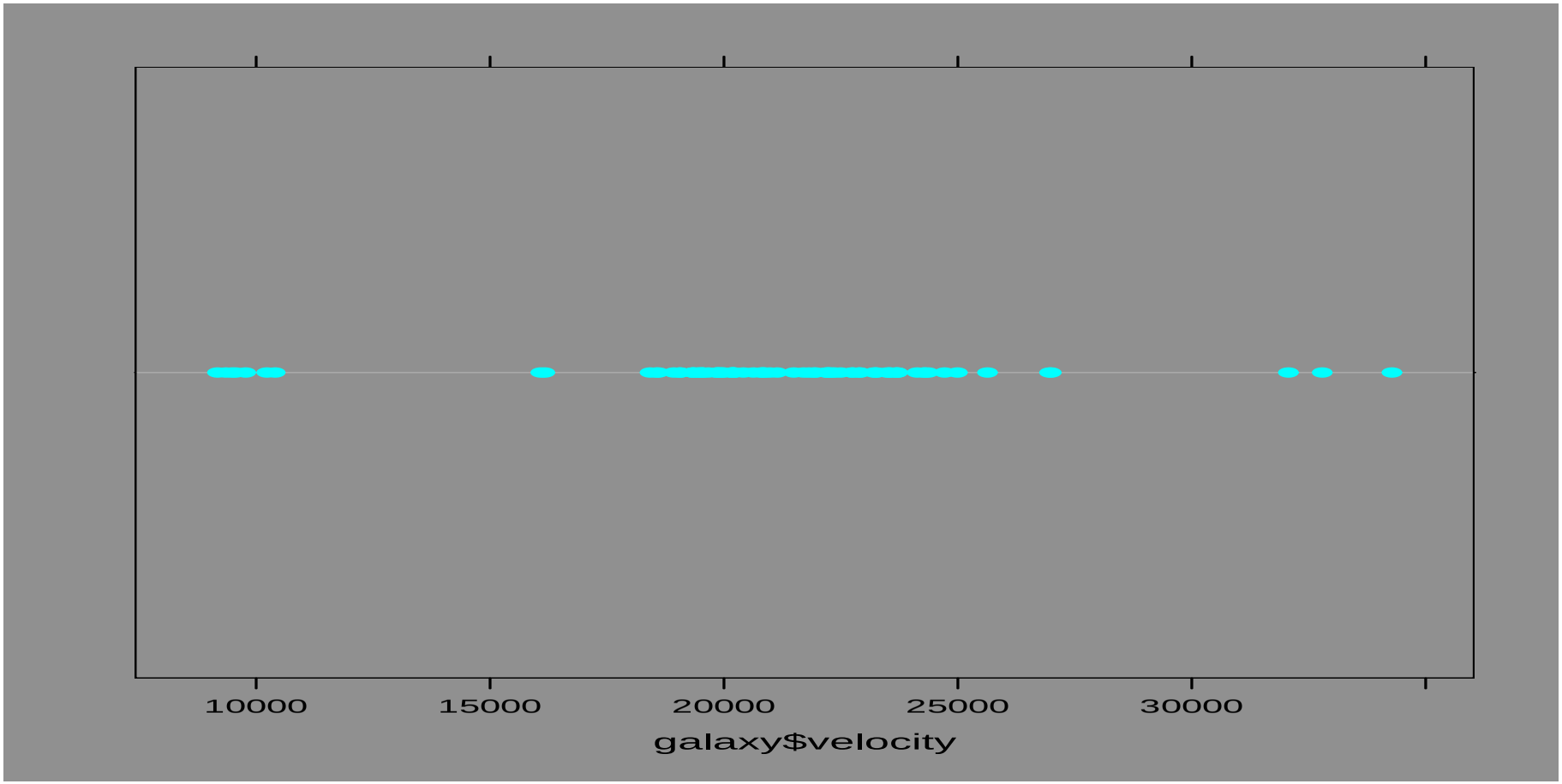
Aim: Estimate

- the mass points z_k and the variances σ_k^2,
- the masses π_k,
- the number of components K

simultaneously.

Example: Galaxy Data

Recession velocities (in km/s) of 82 galaxies.



Finite Gaussian mixture?

Estimation

For fixed K, consider the log-likelihood

$$\ell = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k f(y_i | z_k, \sigma_k^2) \right\},$$

and calculate the score equations

$$\frac{\partial \ell}{\partial z_k} = 0, \quad \frac{\partial \ell - \lambda(\sum \pi_k - 1)}{\partial \pi_k} = 0, \quad \frac{\partial \ell}{\partial \sigma_k} = 0,$$

which turn out to be weighted versions of the single-distribution score equations.

$\Longrightarrow$ can be solved by standard EM algorithm:

Starting points Select starting values z_k^0, π_k^0, and σ_k^0, $k = 1, \ldots, K$.

E-Step Adjust weights given current parameter estimates.

M-Step Update parameter estimates.

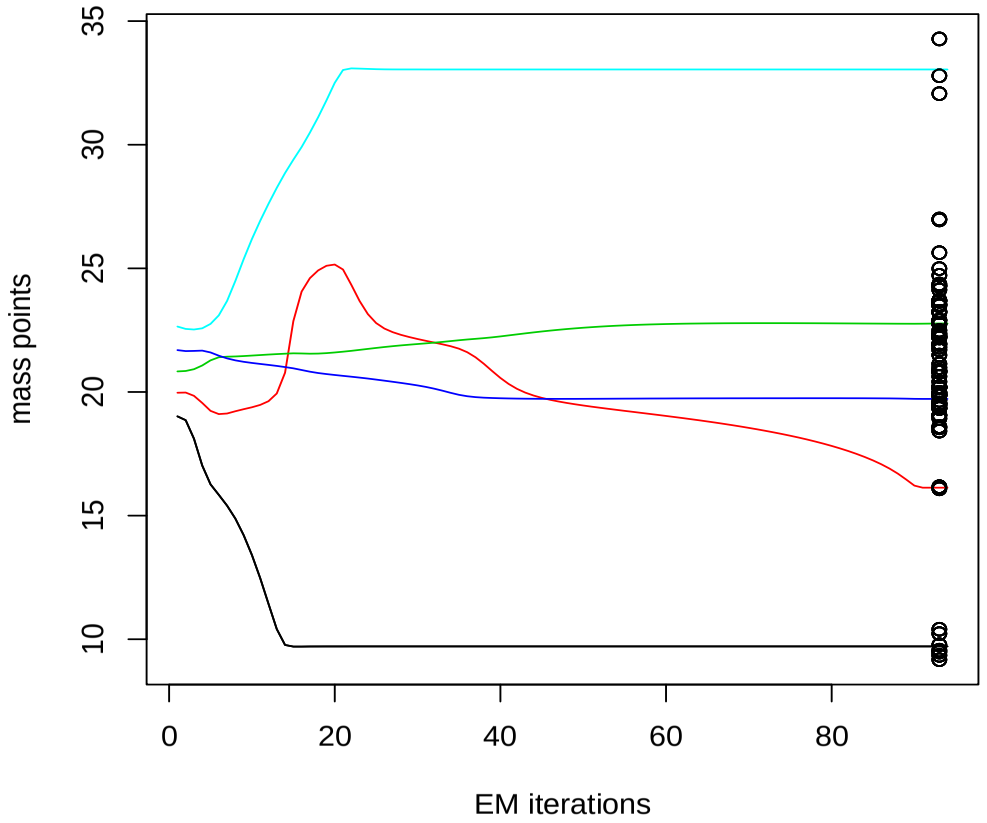
Application on Galaxy Data

Set e.g. K=5:

```
Coefficients:
 MASS1  MASS2  MASS3  MASS4  MASS5
 9.71   16.13  22.78  19.72  33.04
Mixture proportions:
 0.085  0.024  0.512  0.342  0.037
Standard deviations:
 0.423  0.043  1.721  0.626  0.922

-2 log L:     380.9
```

EM Trajectories:



Properties of current NPML implementations (as in GLIM 4)

- The EM algorithm converges in every case.
- NPML estimates are *"impressively stable"* (Aitkin, 1996) and reproducable.
- Results depend heavily on the choice of starting points z_k^0, usually defined as

$$z_k^0 = \bar{y} + \texttt{tol} * \hat{\sigma} * g_k$$

  where `tol`: scaling parameter, g_k: Gauss-Hermite mass points, $\hat{\sigma} = \frac{1}{n}\sum(y_i - \bar{y})^2$.
- Finding the optimal solution requires a tedious grid search for `tol`.
- The EM trajectories behave quite erratically in the first cycles, and tend to cross.
- The positions of the 'optimal' starting points apparently 'have nothing to do' with the optimal mass points. This makes automatic starting point selection difficult.

- General unsolved problem:
  - Set of parameters to estimate:

$$\{\mathsf{K},\ z_1, \ldots, z_K, \pi_1, \ldots, \pi_K, \sigma_1, \ldots, \sigma_k\}.$$

  - Set of tuning parameters to specify beforehand:

$$\{\ \mathsf{K}, z_1^0, \ldots, z_K^0, \pi_1^0, \ldots, \pi_{K-1}^0, \texttt{tol}\}.$$

"one of the things you do not know is the number of things that you do not know"

(Richardson & Green, 1997).

There does not exist any automatic routine to estimate K. Usually, it is increased successively until the likelihood ceases to fall.

Possible improvements

First Step: Damping the EM algorithm.

Shrink estimated standard deviation $\hat{\sigma}_k$ of the mixture components in the $j-th$ cycle

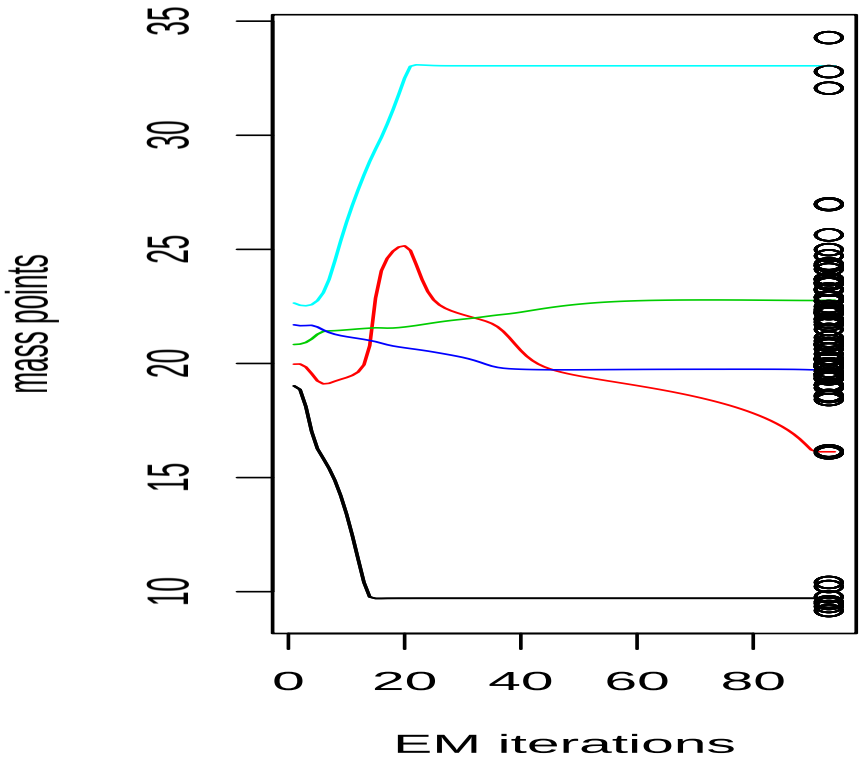
by the factor

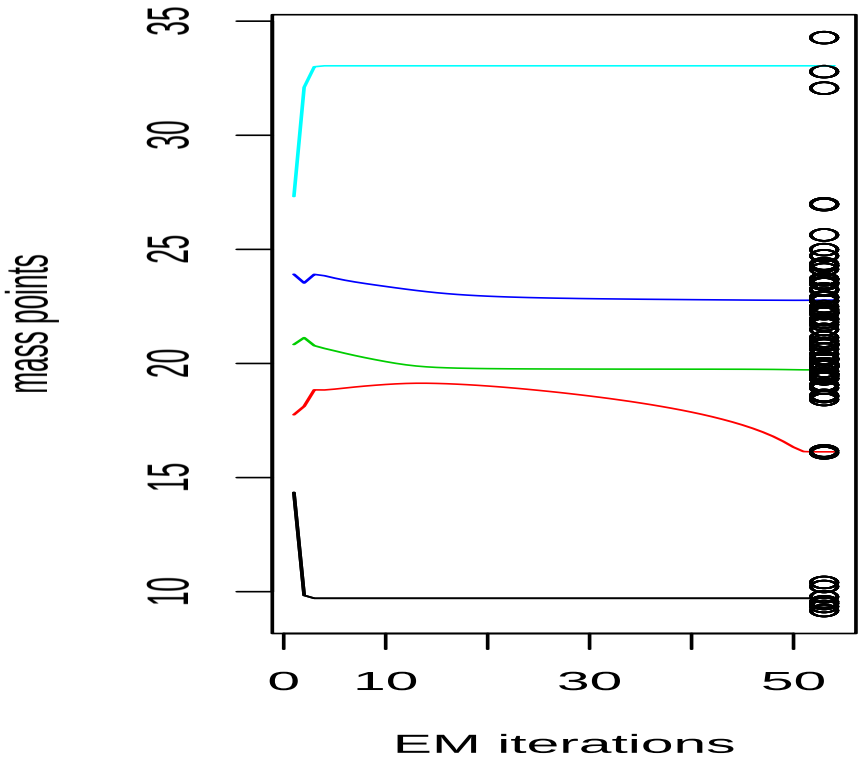
$$d_j = 1 - (1 - \texttt{tol})^j, \qquad (0 < \texttt{tol} \leq 1)$$

i.e. $d_1 = \texttt{tol}$ and $d_j \longrightarrow 1$ for $j \longrightarrow \infty$.

- Damping has main effect in the first cycles.
- Reduces fluctuations and dependence on `tol`.
- Optimal mass points are optimal starting points.

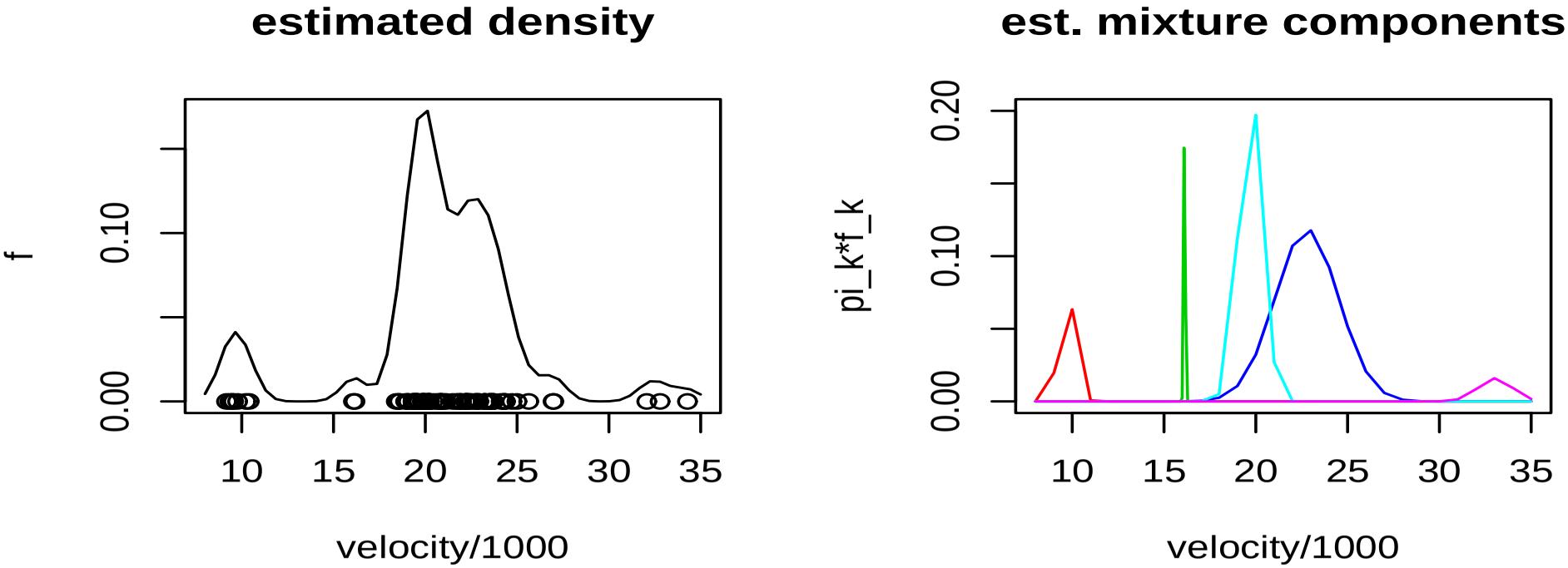
Without damping		With damping
380.9	$-2\log L$	380.9
0.133-0.143	opt. `tol` range	0.45-0.82
93	No. iterations	61





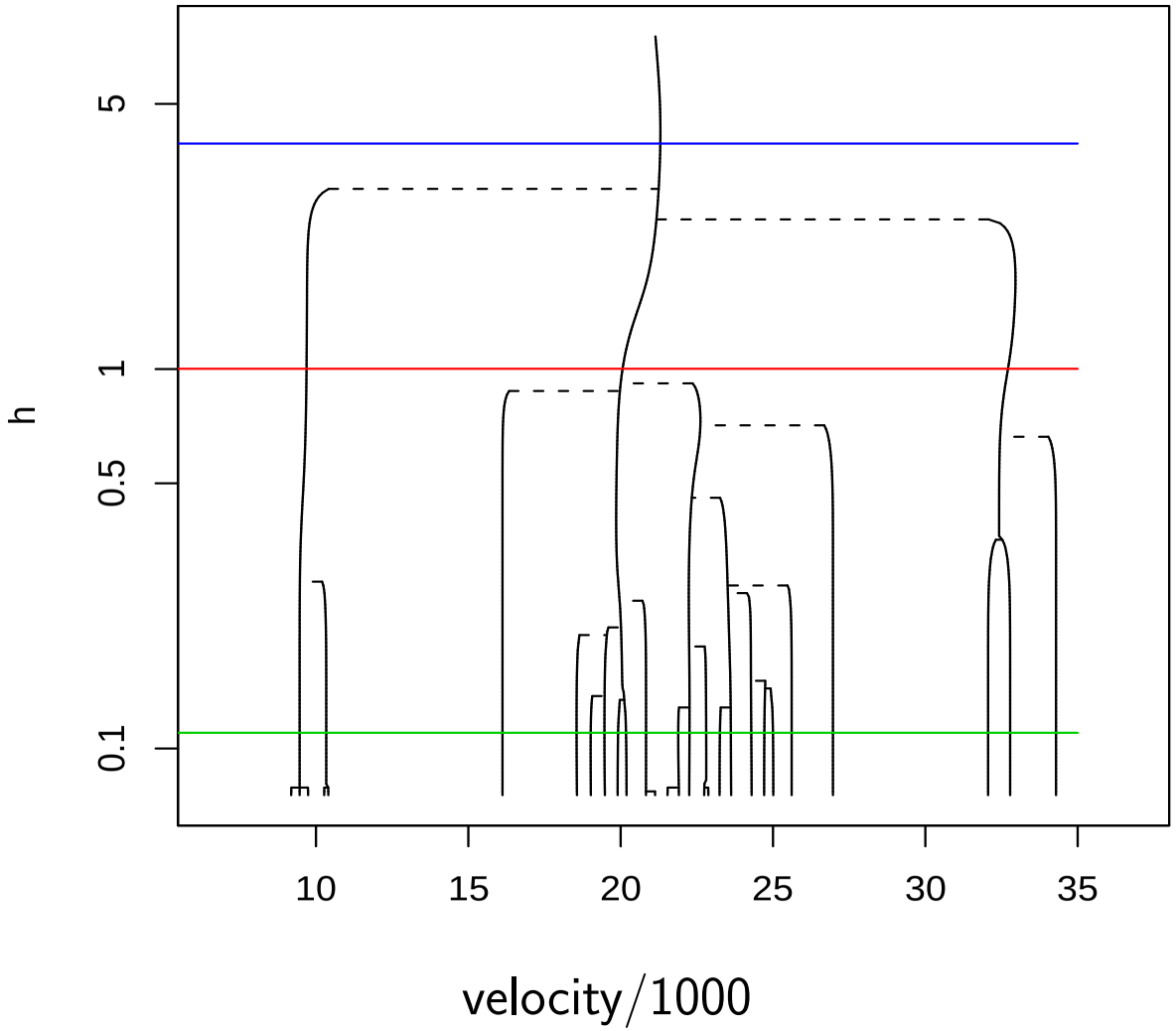
Second step: Find optimal starting points

Idea: Consider density estimate $\hat{f}(y,h) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{Y_i - y}{h}\right)$.



Carreira & Perpiñan, 2003: The number of modes is a *lower bound* for the number of components of a Gaussian mixture. But: Number of modes depends on bandwidth h.

The mode tree (Minnotte & Scott, 1993)



Bandwidth selectors:

BCV (1 mode)

Silverman (3 modes)

AIC (22 modes)

Bandwidth selection in 2 steps

- Calculate Silverman's optimal bandwidth

$$h_{opt} = 0.9An^{-1/2},$$

where $A = \min\{\hat{\sigma}, IQR/1.34\}$

- From that bandwidth, climb down the mode tree until the next critical bandwidth

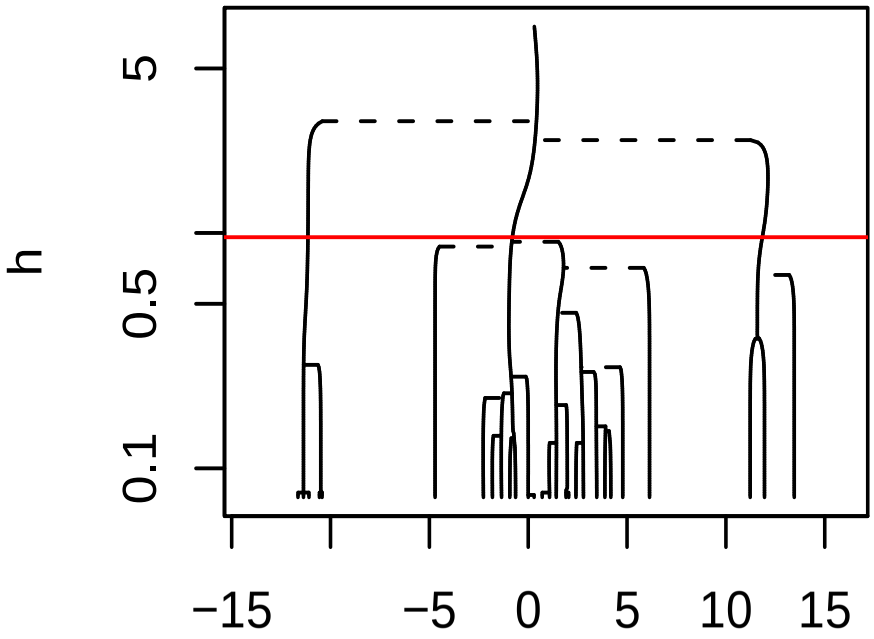
(Silverman, 1981)

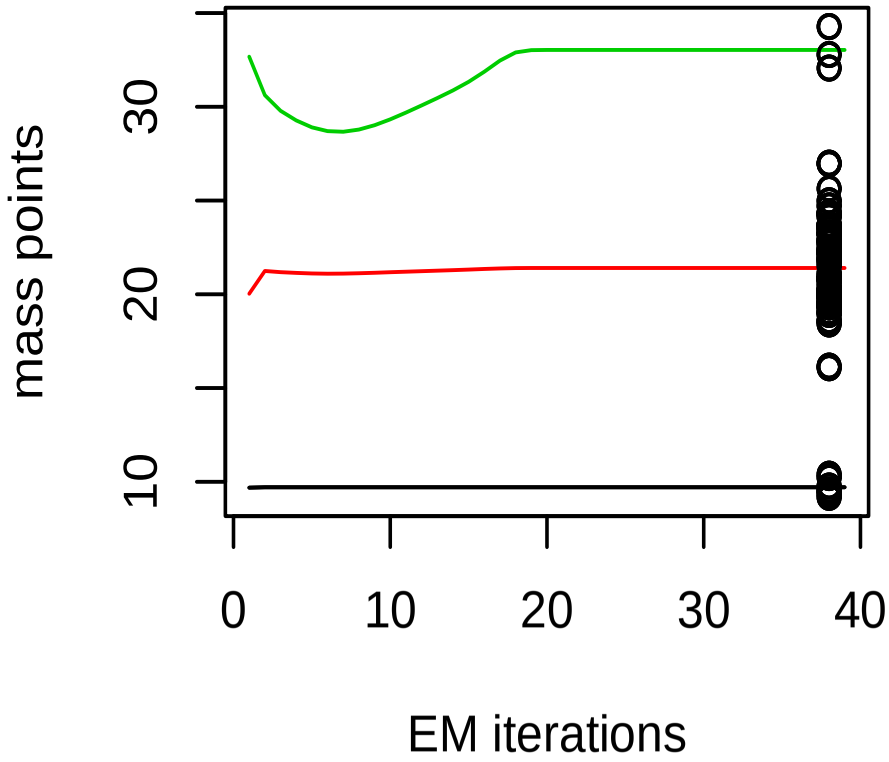
$$h_{crit} = \inf\{h, \hat{f}(\cdot, h) \text{ has at most } k \text{ modes}\}$$

is reached.

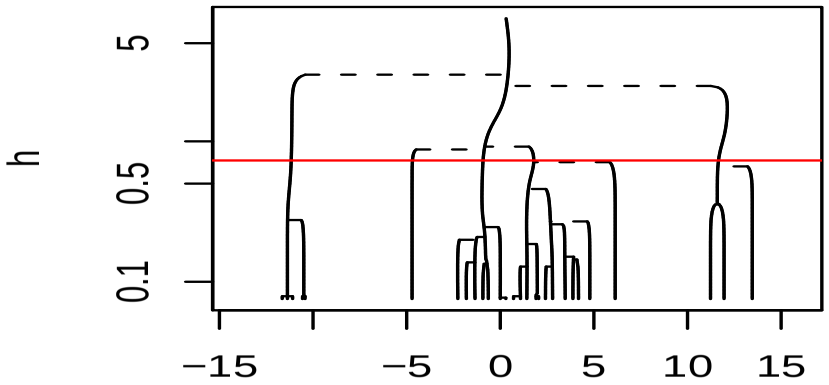
Using h_{crit}, the mode tree gives

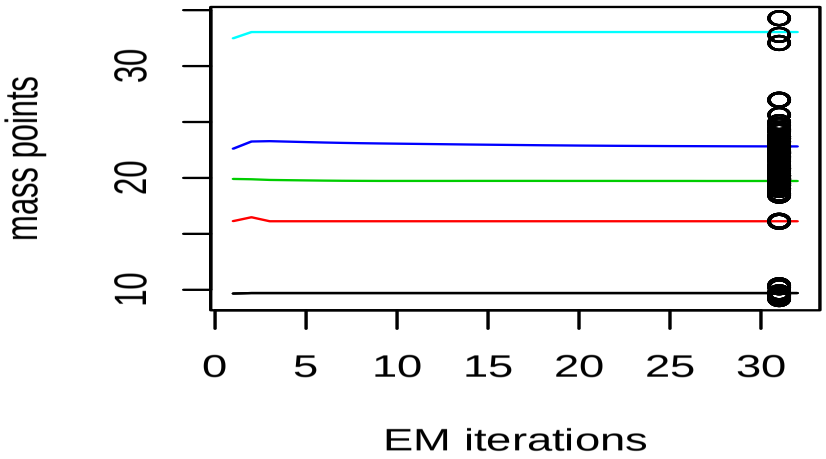
- an estimate for the number of modes, and thus for the number of components
- a very accurate estimate for the location of the mass points, which then can be used as starting points for the EM algorithm.

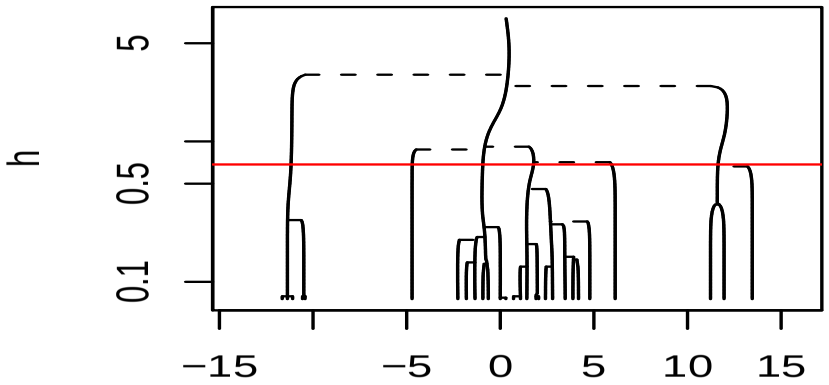


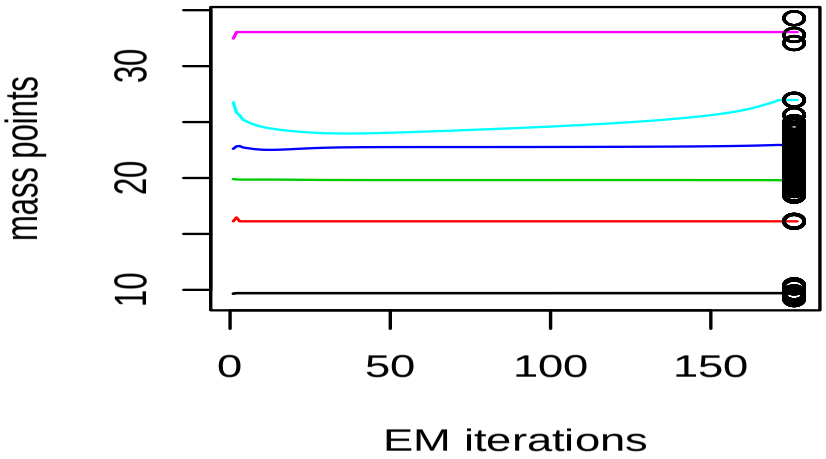


Climbing down the tree









Summary

- Applying a simple damping procedure, the EM algorithm could be stabilzed and made less sensitive to tuning parameters.
- Starting point selection with mode trees works nicely $given\ K$. The starting points are so accurate that one hardly needs EM at all!
- The mode tree is a useful instrument to assess visually the number of components. Applied more general to the 'residuals' $h^{-1}(Y_i) - x_i'\hat{\beta}$ of a GLM, they also work if the multimodal structure cannot be seen in the data cloud itself.
- Mode trees together with a suitable bandwidth selector give a useful $recommenda-tion$ for the choice of K. However, this is no reliable automatic routine, as small variations in the bandwidth may change drastically the number of detected modes.

Everything more general....

- Replace Gaussian by another exponential family distibution
- Set an appropriate link function
- Include explanatory variables
- Random coefficient models
- Variance component models

..... is being implemented in an R package {npml} (Einbeck, Darnell, & Hinde), see

`www.nuigalway.ie/maths/je/npml.html`

References

AITKIN, M. (1996): A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing, 6, 251–262.*

AITKIN, M., FRANCIS, B. and HINDE, J. (2005): *Statistical Modelling in GLIM 4* (Second edition). Oxford, UK.

CARREIRA-PERPIÑAN, M. A. and WILLIAMS, C.K.I. (2003): On the number of modes of a Gaussian mixture. *Lecture Notes in Computer Science, 2695, 625–640.*

LAIRD, N. M. (1978): Nonparametric maximum likelihood estimation of a mixing distribution. *JASA*, **73**, 805–811.

MINNOTTE, M. C. and SCOTT, D. W. (1993): The mode tree: A tool for visualization of nonparametric density features. *JCGS*, **2**, 51-68.

RICHARDSON, S. and GREEN, P. (1997): On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion) *JRSSB*, **59**, 731–792.

SILVERMAN. (1981): Using kernel density estimates to investigate multimodal regression. *JRSSB*, **43**, 97–99.