On Statistical Modelling with Random Effects using Mixture Models

Jochen Einbeck

National University of Ireland, Galway

München, 29. November 2005

joint work with John Hinde, National University of Ireland, Galway.

Work supported by

science foundation ireland
fondúireacht eolaíochta éireann

Outline

- Generalized linear models with random effects,

  Nonparametric maximum likelihood (NPML) estimation

- Example for NPML estimation: Irish suicide data

- Problems of NPML;

  Methodological improvements

- Outlook: Mode Trees

Generalized linear model with random effect

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

with $y_i|z_i$ exponential family distributed. The random effect z_i with distribution $g(\cdot)$

accounts for

- unobserved covariates
- model misspecification
- individual unit variability

⋮

Parameter estimation requires maximizing the marginal likelihood

$$L(\beta, g(z)) = \prod_{i=1}^{n} \int f(y_i|z_i, \beta) g(z_i)\, dz_i.$$

Integral often intractable!

Lots of approaches to solve this problem

(1) Marginal models $\longrightarrow$ GEE

(2) Conditional models $\longrightarrow$ 3 families:

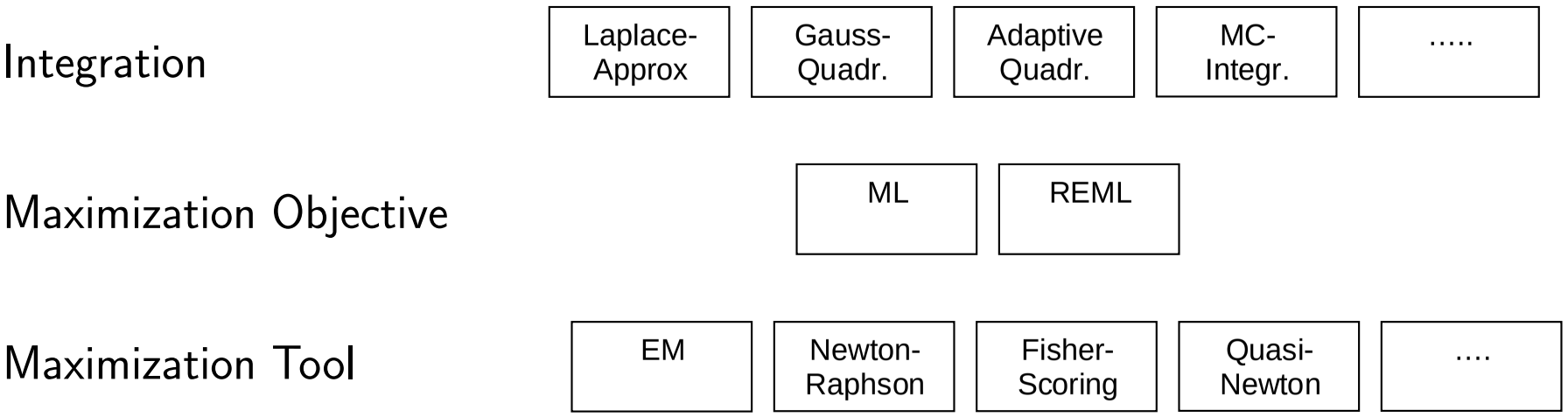
Consider		Fixed part	
	as	fixed	random
Random part	fixed	JML, CML	– –
	random	all other	Bayes, MCMC

Random effect distr.: normal | conjugate | unspecified

- normal: classical; glmm - literature
- conjugate: gives analytical solution, works only for special comb's of Z and $Y|Z$.
- unspecified: NPML

Overview: Estimating glmm's with normal random effects

There exist approaches to nearly all possible combinations of the following 3 levels:



See also McCulloch & Searle (2001), Reithinger (2003), Skrondal & Rabe-Hesketh (2004).

Nonparametric maximum likelihood (NPML, Aitkin, 1996).

Idea: Approximate random effect Z by a finite discrete distribution:

K mass points $\{z_k\}$ with masses $\{\pi_k\}$

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

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

with mass points z_k and masses π_k.

- No parametric assumption about the random effect distribution $g(\cdot)$.
- Simple simultaneous estimation of β, z_k and π_k via EM algorithm.
- Fitted model is a K component mixture model.

Estimation

For a fixed number of mass points K, consider the log-likelihood

$$\ell = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k f(y_i | z_k, \beta) \right\} \equiv \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k f_{ik} \right\}.$$

The score equations for β and z_k turn out to be weighted versions of the single-distribution score equations, with weights

$$w_{ik} = \frac{\pi_k f_{ik}}{\sum_\ell \pi_\ell f_{i\ell}} = P(\text{obs. } y_i \text{ comes from comp. k}).$$

The score equation for π_k gives the simple solution $\hat{\pi}_k = \frac{1}{n} \sum_i w_{ik}$.

$\Longrightarrow$ can be solved by an standard EM algorithm.

EM algorithm for NPML estimation

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

E-Step Adjust weights w_{ik} given current parameter estimates.

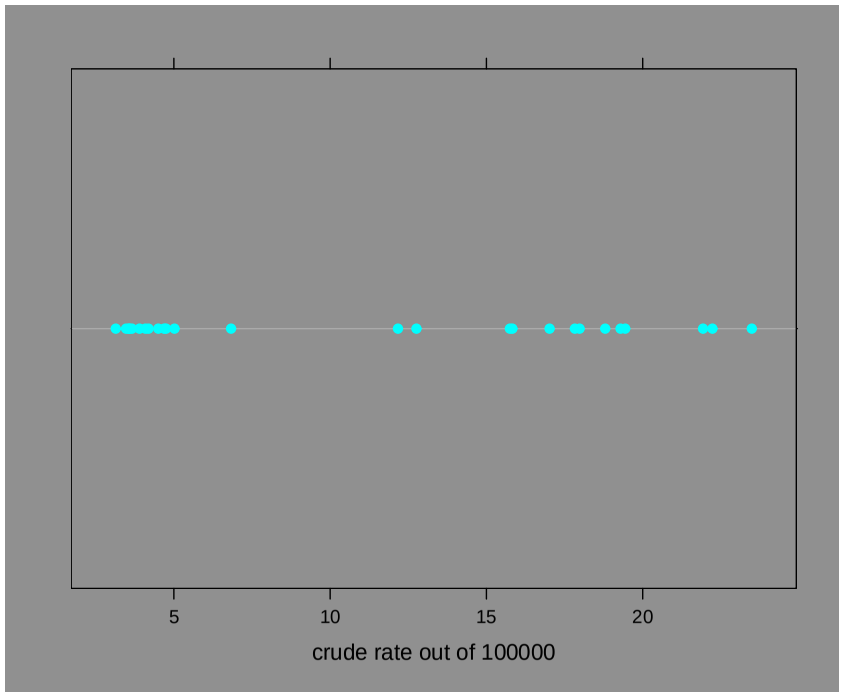
M-Step Update parameter estimates fitting a weighted GLM, with weights w_{ik}.

Example: Irish Suicide rates

- 13 'health regions' (8 health boards + Cork, Dublin, Galway, Limerick, Waterford)
- For each region, we have a total count of suicides over the 10 years, and a corresponding 'crude death rate' out of a population of 100000.
- Explanatory variables: sex, age

Region(s)	Gender	deaths	population	crude death rate
Cork	Female	45	65925	6.83
Cork	Male	144	61298	23.49
Dublin	Female	127	253118	5.02
Dublin	Male	358	227372	15.75
Galway	Female	10	28805	3.47
Galway	Male	41	25897	15.83
⋮				
SHB % Cork	Female	97	204327	4.75
SHB % Cork	Male	413	212499	19.44
WHB % Galway	Female	56	143648	3.9
WHB % Galway	Male	29	150303	19.29

Crude rates :



- Crude rates: based on small observed counts, very variable
- Overall rate hides differences of interest

} need something in between

Modelling suicide rates

We model the number Y of suicides out of a population of size m by a binomial distribution $Y \sim B(m, p)$, with rate p given by

$$\log\left(\frac{p}{1-p}\right) = \text{'RegionalEffect'} + \beta \cdot sex + \ldots$$

- Fixed effect models: 'RegionalEffect' $= \sum_r \alpha_r I_r$

  I_r regional indicator, α_r: parameter for each region. Too much parameters!

- Random effect models: Random effect Z at any appropriate level:

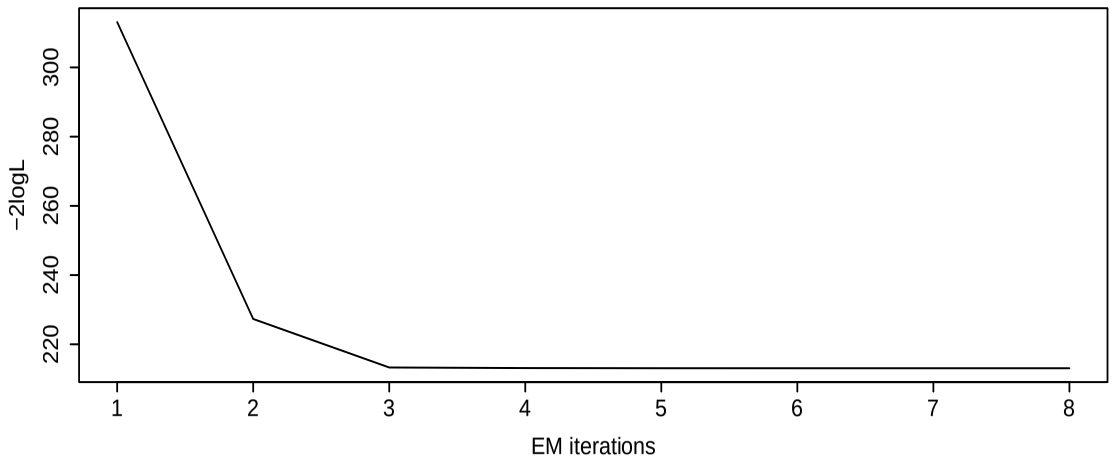
  – observation $\Longrightarrow$ overdispersion

  – region $\Longrightarrow$ regional heterogeneity

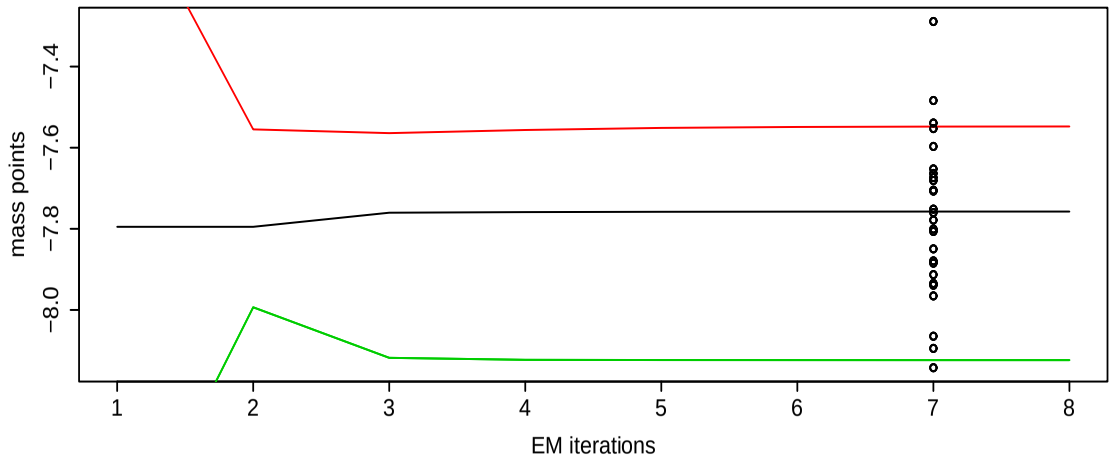
  ("Variance component model", "Two-level-model")

Variance component model for regional random effect: NPML estimation

Disparity trend and EM Trajectories of mass points z_k :




```
Coefficients:
  sex    MASS1   MASS2   MASS3
 1.432  -8.124  -7.757  -7.548
Mixture proportions:
 MASS1   MASS2   MASS3
 0.0996  0.7128  0.1874
-2 log L:      213.1
```



The 3 mass points correspond to regions with high, medium, and low suicide rates.

From crude to modelled ("shrunk") rates

Posterior probabilities w_{ik}

z_1	z_2	z_3	Region
0.00	0.00	1.00	Cork
0.00	1.00	0.00	Dublin
0.06	0.92	0.01	Galway
0.00	0.62	0.38	Limerick
0.23	0.76	0.01	Waterford
1.00	0.00	0.00	EHB % Dublin
0.00	1.00	0.00	Mid WHB % Limerick
0.00	1.00	0.00	Midland HB
0.00	1.00	0.00	NEHB
0.00	1.00	0.00	NWHB
0.00	0.01	0.99	SEHB % Waterford
0.00	0.97	0.03	SHB % Cork
0.00	1.00	0.00	WHB % Galway

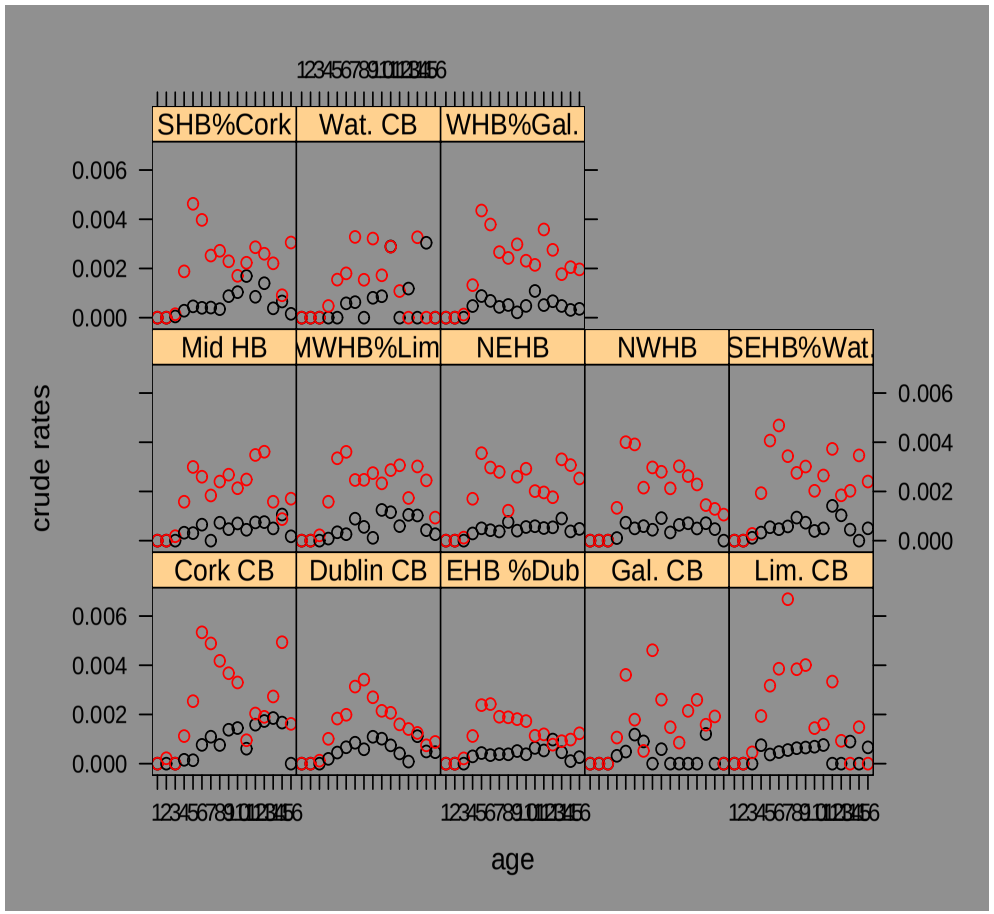
Emp. Bayes $\longrightarrow$

'Suicide league table' for men

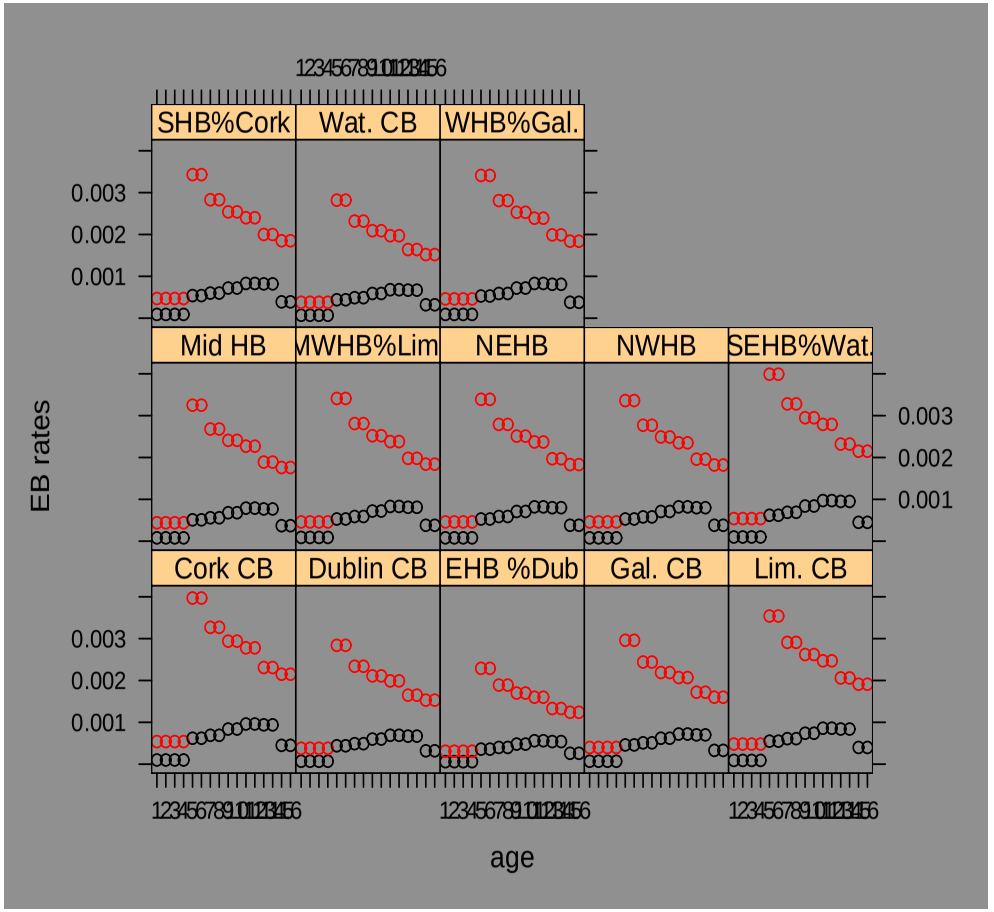
Shrunk Rates	Crude Rates	Region
12.43	12.17	EHB % Dublin
14.83	12.76	Waterford
16.01	15.75	Dublin
17.11	15.83	Galway
17.59	17.02	Midland HB
18.23	17.83	NWHB
18.42	17.98	NEHB
18.59	18.80	Mid WHB % Limerick
18.64	19.44	SHB % Cork
18.66	19.29	WHB % Galway
19.81	22.23	Limerick
21.78	23.49	Cork CB
22.08	21.93	SEHB % Waterford

Inclusion of age (and interaction sex/age)

Crude rates over regions



Modelled ("shrunk") rates over regions



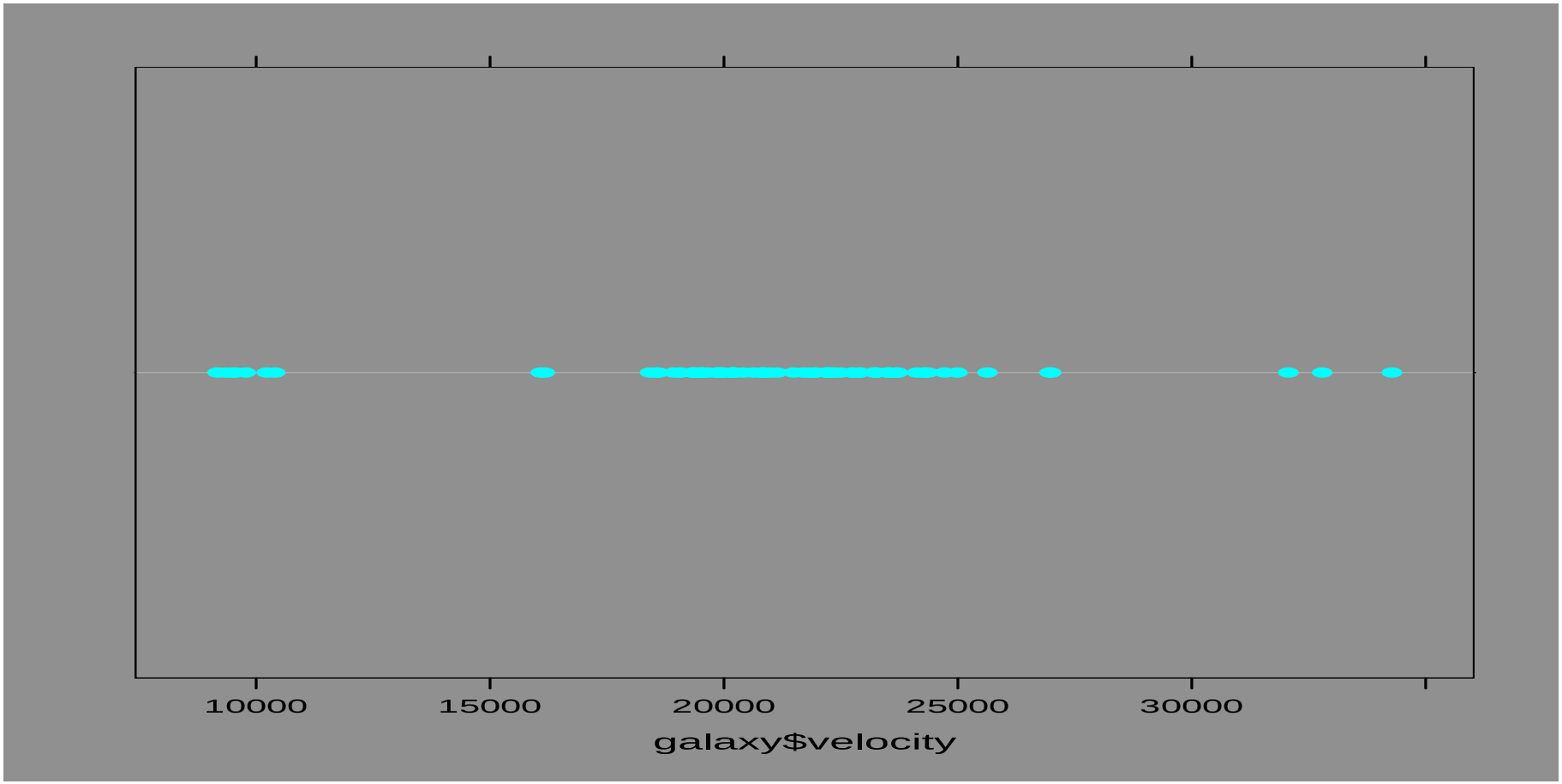
● men ● women

Summary

- Suicide rates are highest in City Cork and SEHB without Waterford, and lowest in region Dublin.
- The modelled ("shrunk") suicide rates of smaller districts (in particular cities Cork, Waterford) are more reliable for the use in a league table than the crude rates.
- Suicide rates tend to be bigger for men than for women, but increase for women and decrease for men with increasing age.
- NPML gives nicely interpretable results (Posterior probabilities, ...) beyond the pure parameter estimates.

Example: Galaxy Data

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



Finite Gaussian mixture?

Fitting a finite Gaussian mixture

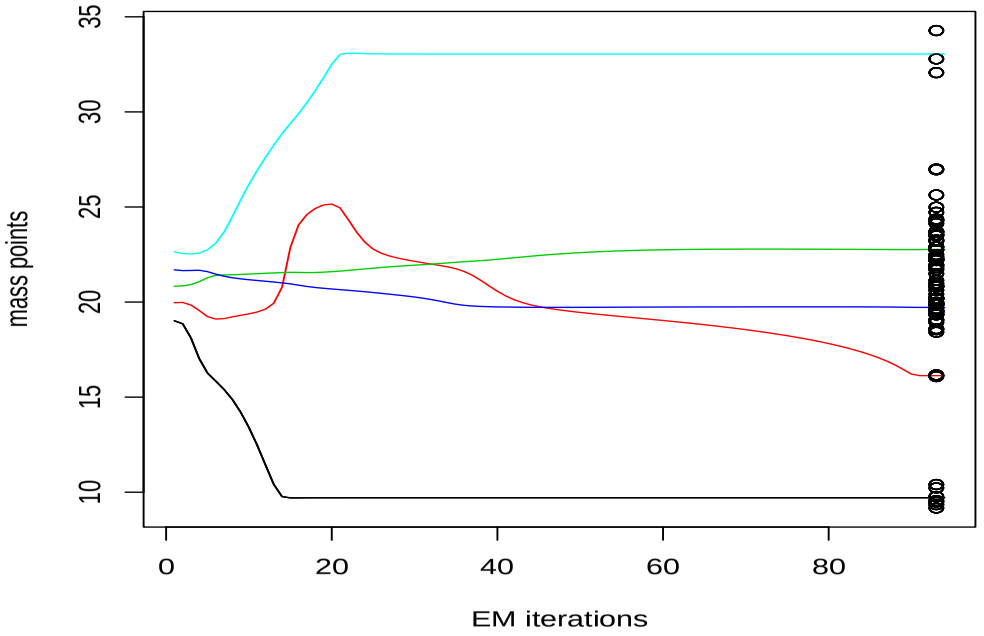
log-likelihood for fixed K

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

E.g. K=5, unequal variances:

```
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:



Problems of current NPML implementations (as in GLIM 4)

- 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 problem: There does not exist an automatic routine to select K.

"one of the things you do not know is the number of things that you do not know" (Richardson & Green, 1997)

Improvement: Damping the EM algorithm

Shrink estimated standard deviation $\hat{\sigma}_{(k)}$ of the mixture components in the $j-th$ cycle

by the factor

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

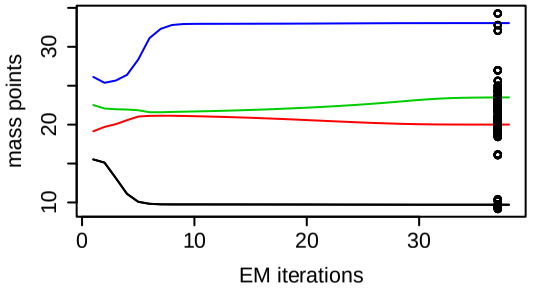
i.e. $d_1 = \texttt{tol}$ and $d_j \longrightarrow 1$ for $j \longrightarrow \infty$.

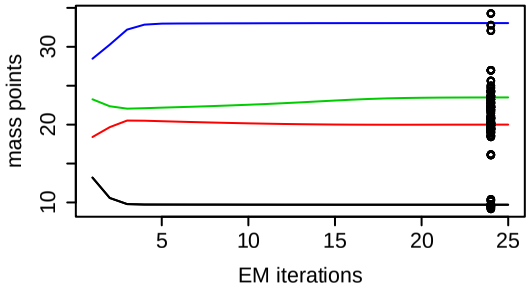
- Damping has main effect in the first cycles.
- Reduces fluctuations and dependence on `tol`.
- Starting in an 'optimal' solution, the damped version does not escape.

EM Trajectories for galaxy data (equal variances)

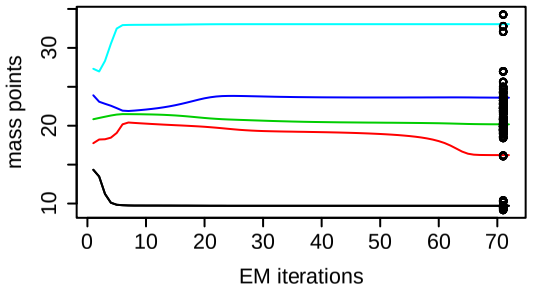
undamped damped

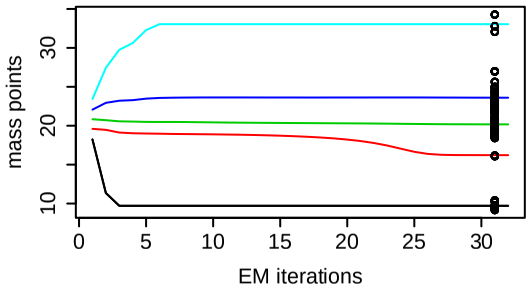
K=4



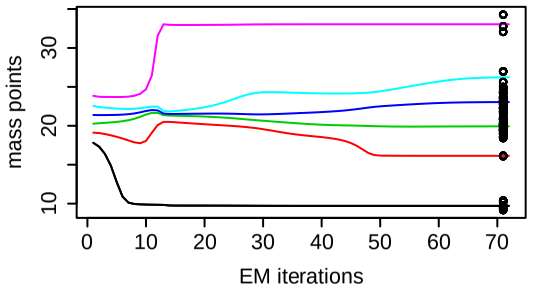


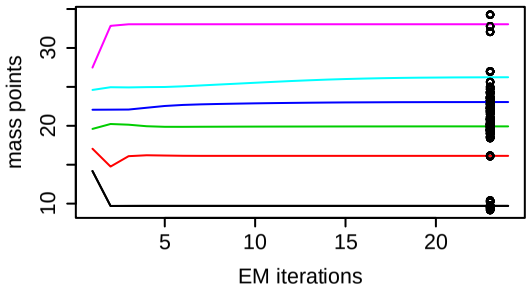
K=5





K=6

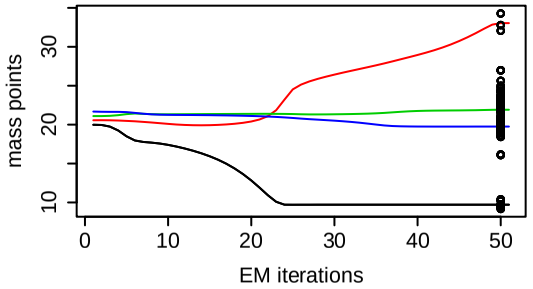


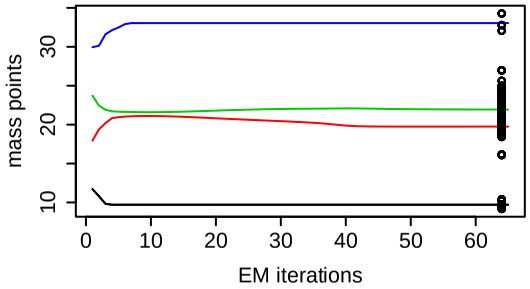


EM Trajectories for galaxy data (unequal variances)

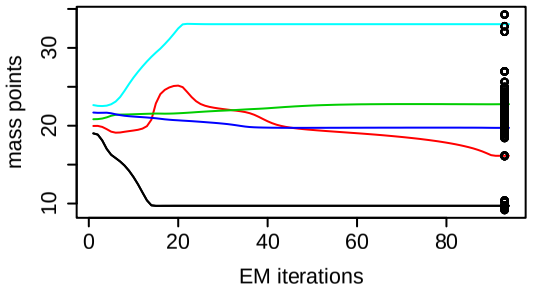
undamped damped

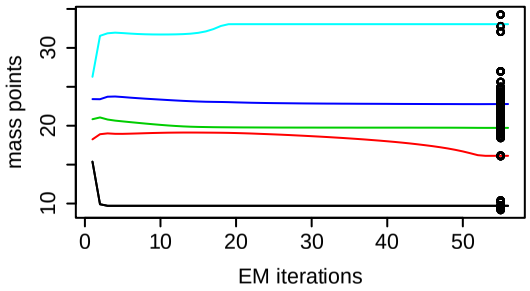
K=4



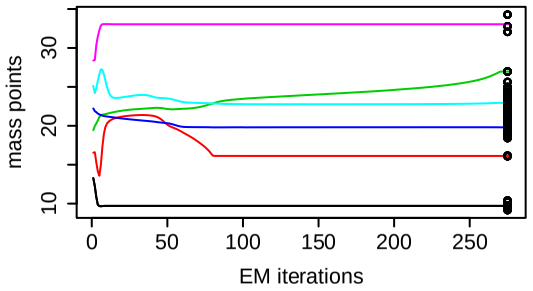


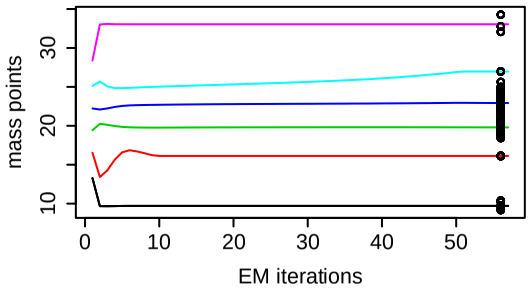
K=5





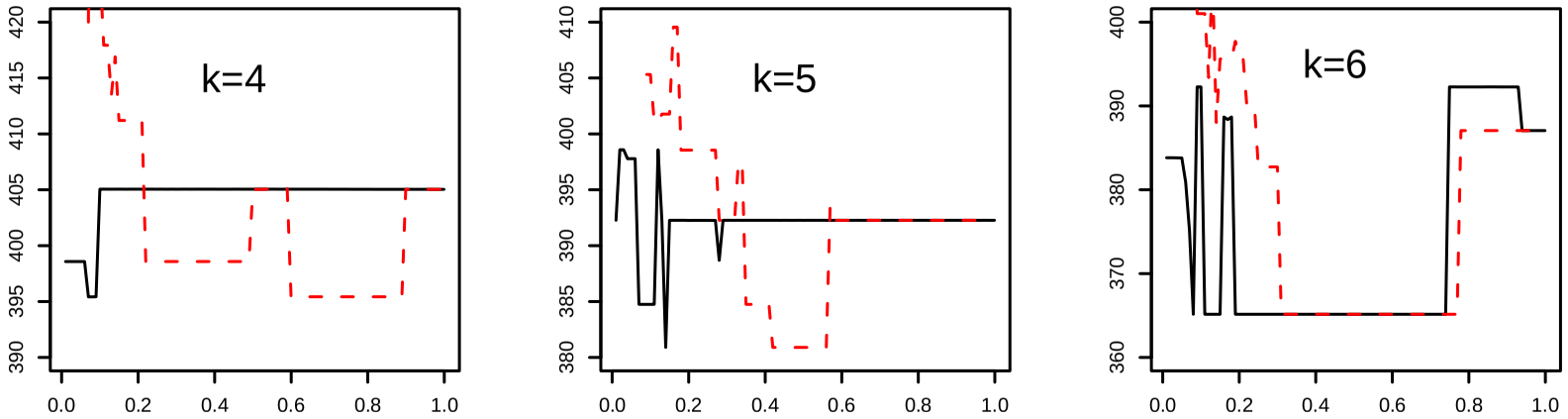
K=6





Sensitivity to tuning parameter reduced

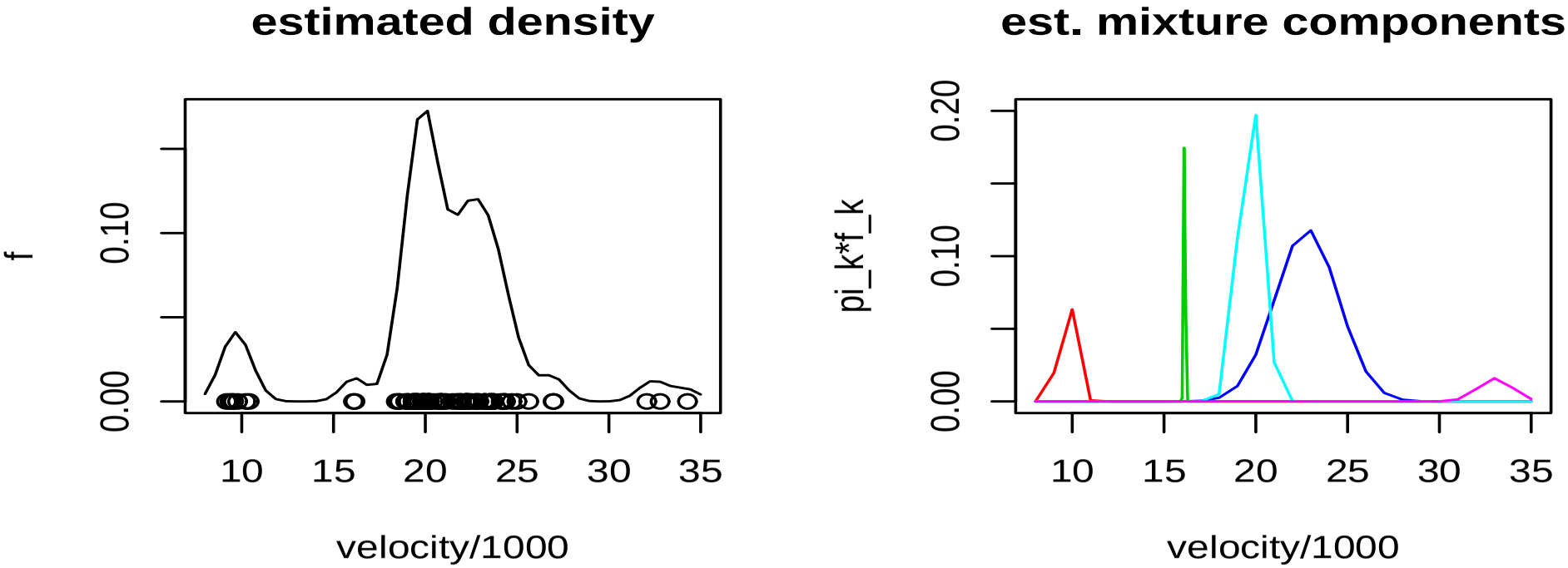
Disparity (i.e. $-2 \log L$) in dependence of `tol`, damped (- -) and undamped (–):



- Damping gives a significant improvement in stability and performance of NPML estimation
- Damping is straightforwardly adapted to other exponential families with dispersion parameter, e.g. Gamma distribution (Einbeck & Hinde, 2005, Austrian Journal of Statistics).

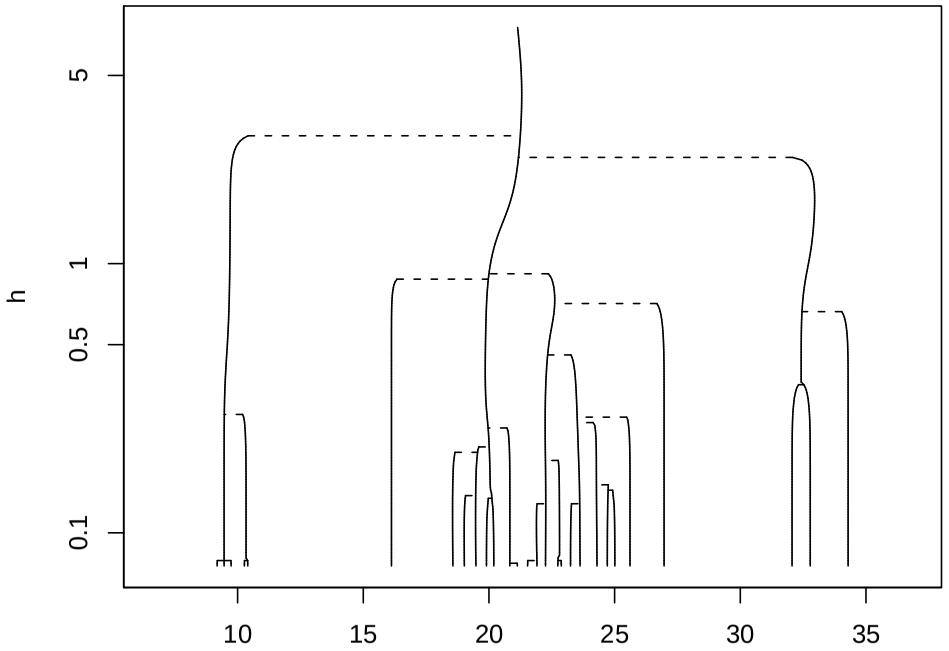
Finding optimal starting points

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



Carreira & Williams, 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)



y_i (velocity/1000)

$\longleftarrow$ fixed effect model $(K = 1)$

$\vdots$

$\longleftarrow$ random effect models $(1 < K < 82)$

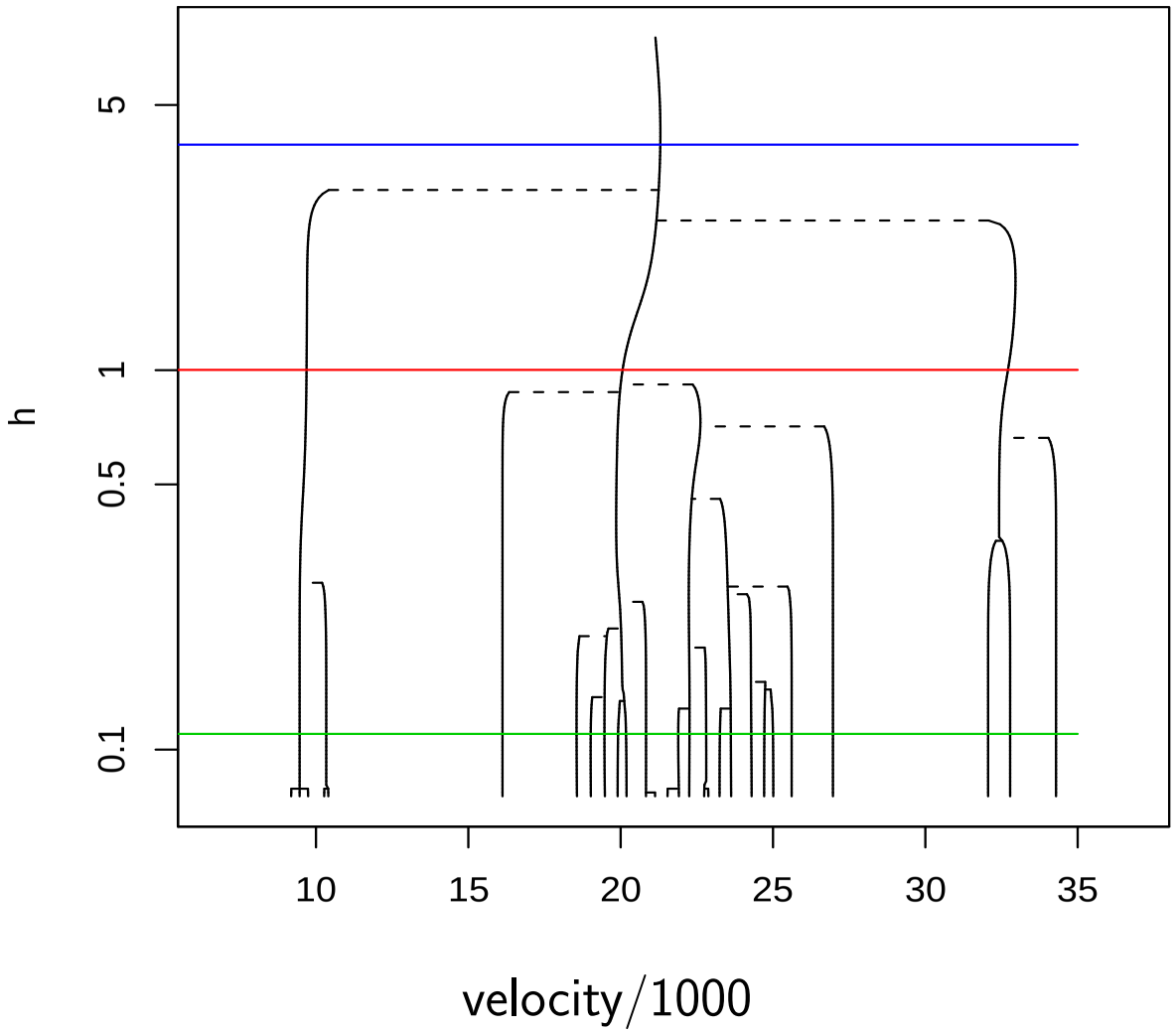
$\vdots$

$\swarrow$ saturated model $(K = 82)$

More generally, applied on the 'residuals' $h^{-1}(y_i) - x_i'\hat{\beta}$ of a GLM:

"zoom into the random effect distribution"

Examples for bandwidth selection



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

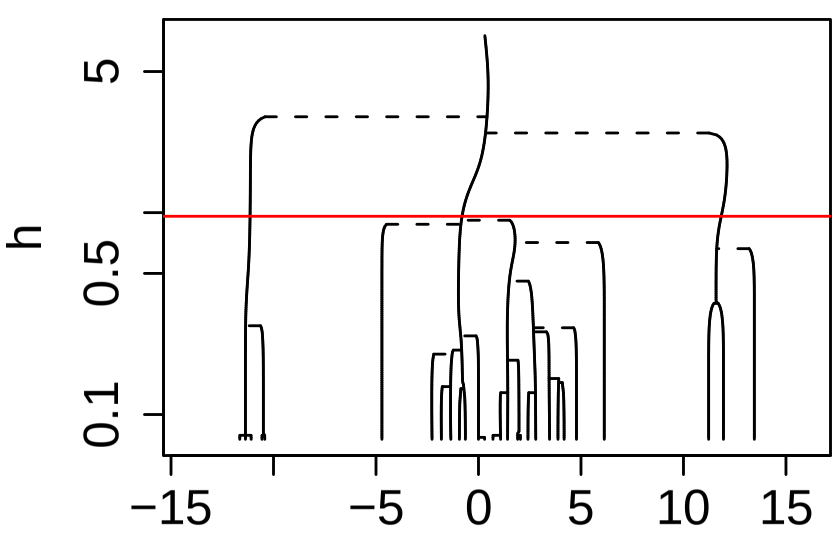
(Silverman, 1981)

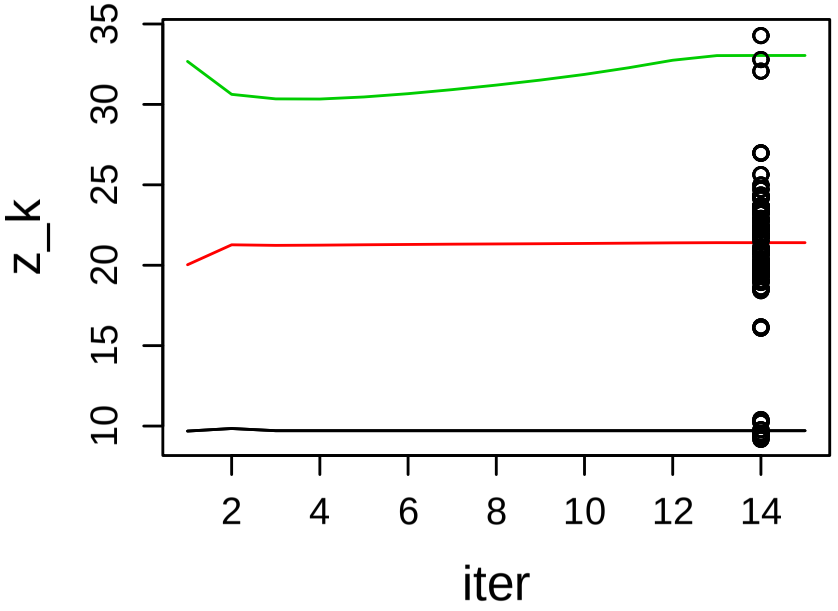
$$h_{crit} = \inf\{h, \hat{f}(\cdot, h) \text{ has at most } k \text{ modes}\}$$

is reached.

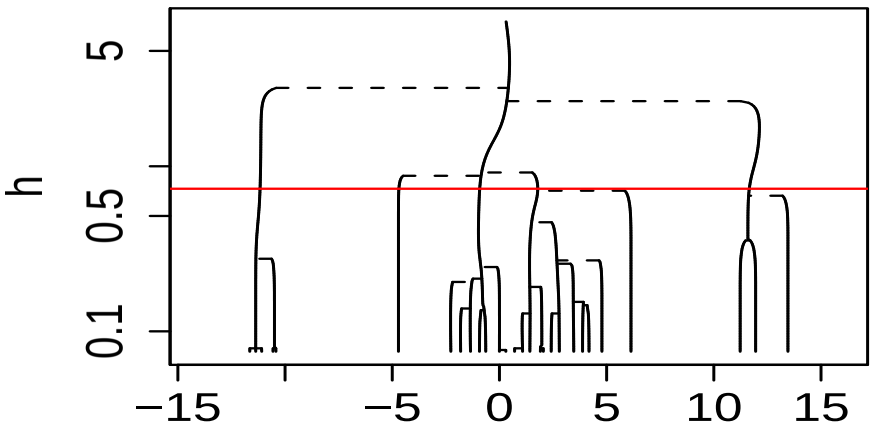
Using h_{crit}, the mode tree gives

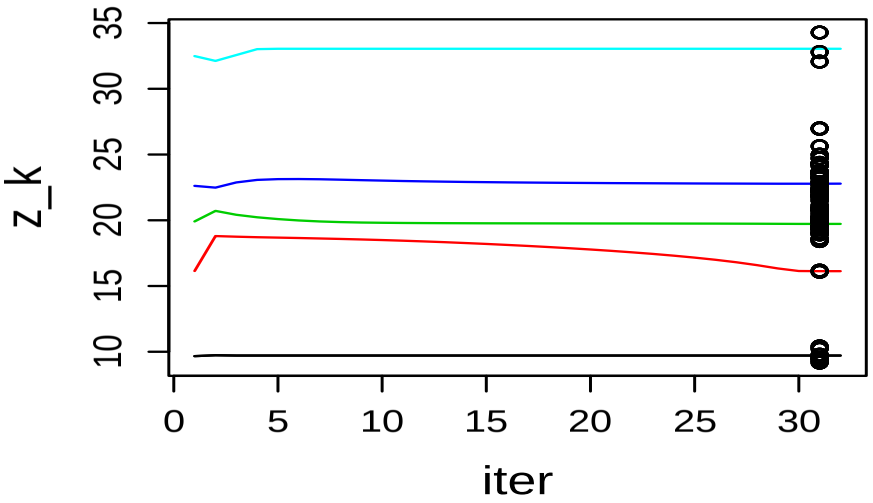
- an estimate for the number of modes, and thus for the number of components
- a very accurate estimate for the location of the mass points, which then can be used as starting points for the EM algorithm. In many cases, these starting points are so accurate that one hardly needs EM at all!

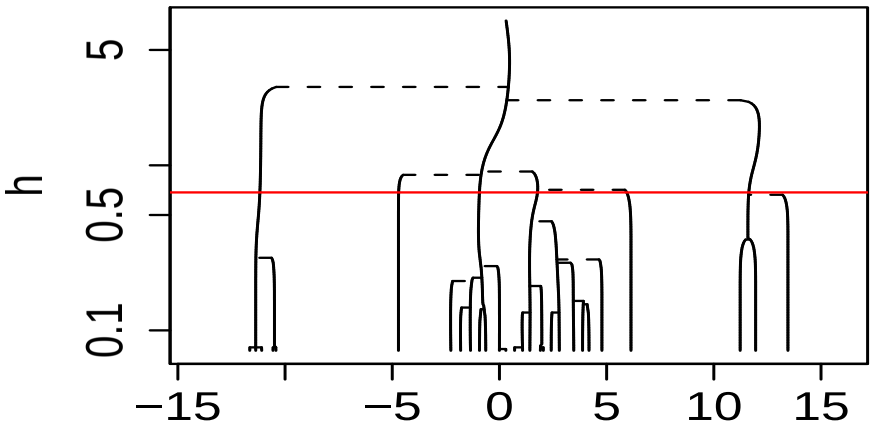


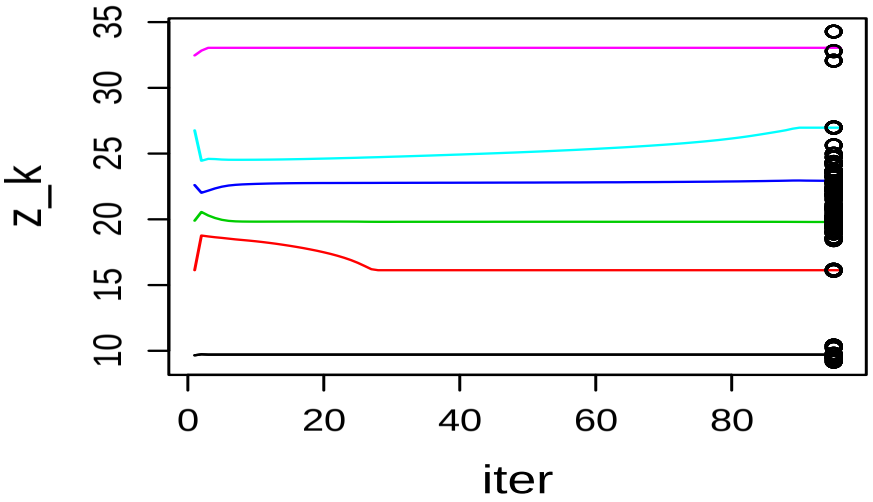


Climbing down the tree









General R Package {npml} (under construction) at

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

in joint work with J. Hinde, based on initial work by R. Darnell.

- Normal, Binomial, Poisson, Gamma - distributed response
- NPML and Gaussian Quadrature
- Variance component models
- Random coefficient models

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.*

EINBECK, J. and HINDE, J. (2005) A note on NPML estimation for exponential family regression models with unspecified dispersion parameter. *Austrian Journal of Statistics*, to appear.

HINDE, J. (1982): Compound Poisson regression models. *Lecture Notes in Statistics* **14** ,109-121.

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

McCULLOCH, C. E. and SEARLE, R. (2001) *Generalized, linear, and mixed models*. New York: Wiley.

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

REITHINGER, F. (2003): *Generalized linear models with random effects and smooth compoments.* Diplomarbeit, LMU München.

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.

SKRONDAL, A. and RABE-HESKETH, S. (2004): *Generalized latent variable modelling*. Boca Raton: Chapman & Hall/CRC.