AUSTRIAN JOURNAL OF STATISTICS
Volume 35 (2006), Number 2&3, 233–243

# A Note on NPML Estimation for Exponential Family Regression Models with Unspecified Dispersion Parameter

Jochen Einbeck and John Hinde
National University of Ireland, Galway, Ireland

**Abstract:** Nonparametric maximum likelihood (NPML) estimation for exponential families with unspecified dispersion parameter $\phi$ suffers from computational instability, which can lead to highly fluctuating EM trajectories and suboptimal solutions, in particular when $\phi$ is allowed to vary over mixture components. In this paper, a damped version of the EM algorithm is proposed to cope with these problems.

**Keywords:** EM Algorithm, Random Effect Models, Nonparametric Maximum Likelihood, Overdispersion, Gamma Distribution, Generalized Linear Model.

## 1 Introduction

Random effects have become a standard concept in statistical modelling in the last decades. One of the most important model classes for the use of random effects is the generalized linear model. Assume there is given a set of explanatory vectors $x_1, \ldots, x_n$ and a set of observations $y_1, \ldots, y_n$ sampled from an exponential family distribution $f(y_i|\beta, \phi_i)$ with dispersion parameter $\phi_i$. In a generalized linear model (GLM), predictors and response are assumed to be related through a link function $h$,

$$\mu_i \equiv \mathrm{E}(y_i|\beta, \phi_i) = h(\eta_i) \equiv h(x_i'\beta)\,.$$

While the dispersion $\phi_i$ is fixed e.g. for the Poisson or Binomial distributions, it may be considered as an additional model parameter for other exponential family distributions. For instance, in case of a normal distribution $N(\mu, \sigma^2)$, the dispersion is given by $\phi = \sigma^2$. In the case of a gamma-distribution, usually written as $\Gamma(\nu, \nu/\mu)$, with shape $\nu$ and rate $\nu/\mu$, the dispersion takes the form $\phi = 1/\nu$. The variance $\sigma_i^2 = \mathrm{var}(y_i|\beta, \phi_i) = \phi_i v(\mu_i)$ depends on a function $v(\mu_i)$ which is entirely determined by the choice of the particular exponential family. However, often the actual variance in the data is larger than the variance according to this strict mean-variance relationship. Reasons for this effect, called overdispersion, might be correlation in the data or important explanatory variables not included in the model. In order to account for additional unexplained variability of the individual observations, a random effect $z_i$ with density $g(z)$ is included into the linear predictor

$$\eta_i = \beta' x_i + z_i\,.$$

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

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

where $f_{ik} = f(y_i|z_k, \beta, \phi_k)$, $z_k$ are the mass points and $\pi_k$ their masses. In many applications it will be sufficient to work with a constant dispersion $\phi \equiv \phi_k$, $k = 1, \ldots, K$. The log-likelihood

$$\ell = \log L = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k f_{ik} \right\} \tag{2}$$

can be maximized as outlined in Section 2.1. In the special case of a normally distributed random effect, one can employ tabulated Gauss-Hermite integration points and masses for $z_k$ and $\pi_k$ and consider these values as constants (Hinde, 1982). For unspecified $g(\cdot)$, they have to be calculated simultaneously with the model parameters by the EM algorithm. As in this case no parametric specification of the random effect distribution is necessary, one refers to this method as 'Nonparametric Maximum Likelihood' (NPML) estimation, which was adapted to the framework of overdispersed generalized linear models by Aitkin (1996).

Though NPML estimation via EM is generally acknowledged to be "impressively" stable (Aitkin, 1996), several authors have expressed concerns. The main problem is that the marginal log-likelihood is maximized for a *fixed* number of components $K$. However, as Böhning (1999) clearly points out, the log-likelihood is a concave functional on the set of *all* discrete distributions (i.e. leaving $K$ unspecified), but is not concave for fixed $K$. Hence, depending on the choice of the EM starting values, various local maxima may be found, and the EM trajectories may fluctuate highly, as will be demonstrated in Section 2.2. In the `GLIM4` implementation, this problem is addressed by means of an additional scaling parameter `tol` (Aitkin and Francis, 1995), as explained in Section 2.1.

The novel contribution of this paper is twofold. Firstly, in Section 3 we propose a damping procedure for fitting Gaussian mixtures with equal or unequal variances, which stabilizes the EM algorithm, alleviates the problem of EM starting point selection, and improves in some cases the 'optimal' results so far obtained with classical NPML. We use an `R` implementation of Aitkin's (1996) NPML algorithm, which was adapted from the `GLIM4` functions `alldist` and `normvar` (Aitkin and Francis, 1995). Secondly, in Section 4 we transfer the concept to the gamma distribution. EM-based NPML estimation for gamma-distributed response had apparently never been implemented before, possibly due to the algorithmic problems mentioned above (`GLIM4` and `C.A.MAN` (Böhning et al., 1992) just support the exponential distribution, i.e. $\nu = 1$). The `R` code can be found at `www.nuigalway.ie/maths/je/npml.html`.

# 2 Observing the EM Trajectories

## 2.1 The EM Algorithm for NPML Estimation

From the log-likelihood (2) one gets the score equations

$$\frac{\partial \ell}{\partial \beta} = 0\,, \quad \frac{\partial \ell}{\partial z_k} = 0\,, \quad \frac{\partial \ell}{\partial \phi_k} = 0\,, \qquad k = 1, \ldots, K\,, \tag{3}$$

which turn out to be weighted versions of the single-distribution score equations (Aitkin et al., 2005, p. 84ff, 416ff, and 457ff), with weights

$$w_{ik} = \frac{\pi_k f_{ik}}{\sum_\ell \pi_\ell f_{i\ell}}\,. \tag{4}$$

The weights $w_{ik}$ can be interpreted as posterior probabilities that the observation $y_i$ comes from component $k$. The score equation for the mixture proportions, $[\partial\ell - \lambda(\sum \pi_k - 1)]/\partial\pi_k = 0$, gives the ML estimate $\hat{\pi}_k = \sum_i w_{ik}/n$, which is the average posterior probability for component $k$. The parameters $\beta$, $\phi_k$, $z_k$, and $\pi_k$ can now be simultaneously estimated by the standard EM algorithm:

**Starting points** Select starting values $\beta^{(0)}$, $\phi^{(0)}$, $z_k^{(0)}$, and $\pi_k^{(0)}$, $k = 1, \ldots, K$.

**E-Step** Adjust weights using formula (4) with current parameter estimates.

**M-Step** Update parameter estimates fitting a weighted GLM.

We now describe the first cycle of the algorithm in more detail. Initially, a GLM (without random effect) is fitted, giving initial estimates $\beta^{(0)}$, $\eta_i^{(0)} = x_i'\beta^{(0)}$, and $\phi^{(0)}$. The mass points are usually initialized as

$$z_k^{(0)} = d \cdot \sigma^{(0)} \cdot g_k \tag{5}$$

with Gauss-Hermite quadrature points $g_k$ and masses $\pi_k^{(0)}$, which are tabulated in standard references as e.g. Abramowitz and Stegun (1970). The tuning parameter $d$ corresponds to `tol` in `GLIM4` and scales the starting points outwards ($d > 1$) or inwards ($0 < d < 1$). As there is not much profit in placing starting points beyond the range of the random effect distribution, the spread of which is captured by the initial standard deviation $\sigma^{(0)}$, we restrict this paper to the choice $0 < d \leq 1$. For a Gaussian response distribution, $\sigma^{(0)}$ is simply given by $(\phi^{(0)})^{1/2}$, otherwise it can be calculated from the 'residuals' on the linear predictor scale by $[\sum(h^{-1}(y_i) - x_i'\beta^{(0)})^2/n]^{1/2}$. One obtains the extended linear predictor for the $k-$th component

$$\eta_{ik}^{(0)} = \eta_i^{(0)} + z_k^{(0)}\,. \tag{6}$$

For the rest of this and the following section, we constrain ourselves to the special case of a mixture of normal models $N(\mu_k, \sigma^2_{(k)})$ with dispersion $\phi_{(k)} = \sigma^2_{(k)}$, where the subscript $(k)$ indicates that the component variances may or may not be equal. In either case, we compute an initial value $f_{ik}^{(0)}$ of $f_{ik}$ via

$$f_{ik}^{(0)} = f(y_i | h(\eta_{ik}^{(0)}), \beta^{(0)}, (\sigma^{(0)})^2)\,. \tag{7}$$

From this one gets in an 'initial E-Step' the weights $w_{ik}^{(1)} = \pi_k^{(0)} f_{ik}^{(0)} / \sum_\ell \pi_\ell^{(0)} f_{i\ell}^{(0)}$, and in the subsequent M-Step one obtains the parameter estimates by solving the score equations (3). In practice, this is done by fitting a weighted GLM with expanded design matrix and weights $w_{ik}^{(1)}$ as specified above (The original design matrix $X = (x_1, \ldots, x_n)'$ is replicated $K$ times and the replicates are joined vertically. Then, $K$ columns are added, where each of them has entries '1' for one of the components, and zero otherwise).

From the resulting estimates of this first cycle, say $z_k^{(1)}$, $\beta^{(1)}$, and $\sigma_{(k)}^{(1)}$, one gets an updated value $f_{ik}^{(1)} = f(y_i|h(x_i'\beta^{(1)} + z_k^{(1)}), \beta^{(1)}, (\sigma_{(k)}^{(1)})^2)$. The estimated masses $\pi_k^{(1)} = \sum_i w_{ik}^{(1)}/n$ are then used together with $f_{ik}^{(1)}$ to update the log-likelihood (2) and the weights (4), and so on. This is continued until the change in disparity (i.e. $-2\ell$) between the current and previous cycle falls below a certain threshold, e.g. 0.001.

## 2.2 Illustrative Example: the Galaxy Data

As an example, we re-analyze the galaxy data (Postman et al., 1986), which are the recession velocities, in units of $10^3$km/s, of 82 galaxies receding from our own. The data set, given in increasing order, is given below.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 9.172 | 9.350 | 9.483 | 9.558 | 9.775 | 10.227 | 10.406 | 16.084 | 16.170 | 18.419 |
| 18.552 | 18.600 | 18.927 | 19.052 | 19.070 | 19.330 | 19.343 | 19.349 | 19.440 | 19.473 |
| 19.529 | 19.541 | 19.547 | 19.663 | 19.846 | 19.856 | 19.863 | 19.914 | 19.918 | 19.973 |
| 19.989 | 20.166 | 20.175 | 20.179 | 20.196 | 20.215 | 20.221 | 20.415 | 20.629 | 20.795 |
| 20.821 | 20.846 | 20.875 | 20.986 | 21.137 | 21.492 | 21.701 | 21.814 | 21.921 | 21.960 |
| 22.185 | 22.209 | 22.242 | 22.249 | 22.314 | 22.374 | 22.495 | 22.746 | 22.747 | 22.888 |
| 22.914 | 23.206 | 23.241 | 23.263 | 23.484 | 23.538 | 23.542 | 23.666 | 23.706 | 23.711 |
| 24.129 | 24.285 | 24.289 | 24.366 | 24.717 | 24.990 | 25.633 | 26.960 | 26.995 | 32.065 |
| 32.789 | 34.279 | | | | | | | | |

This is a very simple data situation with a one-dimensional response and no explanatory variables. Writing this as a random effect model, one has a set of observations $y_1, \ldots, y_n$, where $y_i \sim N(z_i, \sigma_i^2)$, $i = 1, \ldots, n$, with $z_i \sim Z$, where $Z$ is left unspecified. The expectation takes the simple form $\mu_i = \mathrm{E}(y_i|z_i, \sigma_i^2) = z_i$ and the marginal mixture density is given by $\sum_{k=1}^K \pi_k f(y|z_k, \sigma_k^2)$, where $f(y|z_k, \sigma_k^2)$ is a normal density with mean $z_k$ and standard deviation $\sigma_k$. For fixed $K$, the mass points $z_k$, the masses $\pi_k$, and the standard deviances $\sigma_k$ can be estimated by NPML. Our particular interest is not the actual parameter estimates, but rather the structure of the 'EM trajectories', obtained by plotting the mass points $z_k^{(j)}$ over the iteration number $j$. In Figure 1, we show the EM trajectories for $K = 4, 5, 6$, for a choice of $d$ minimizing the disparity

$$-2\ell = -2\sum_{i=1}^n \log\left(\sum_{k=1}^K \pi_k f_{ik}\right) = -\frac{n}{2}\log 2\pi + \sum_{i=1}^n \log\left(\sum_{k=1}^K \frac{\pi_k}{\sigma_k}\exp -\frac{(y_i - z_k)^2}{2\sigma_k^2}\right),$$

and for equal and unequal standard deviations $\sigma_k$. The disparity values are provided in the first column of Table 1 and 2, respectively. One observes immediately that

- For equal variances, all trajectories except the outer ones have to pass a very narrow bottleneck in the early phase of the EM algorithm.
- For unequal variances, the trajectories in the early cycles behave erratically and even tend to cross. The bottleneck is also visible, though less pronounced.
- The position of the final mass points seem to have 'nothing to do' with the position of the starting points!
- The number of iterations for convergence can get very large for unequal variances and a larger number of mass points.

Table 1: Disparities for galaxy data (equal variances)

| $k$ | $-2\log L$ | | |
|---|---|---|---|
| Software | R | R | GLIM4 |
| Damping | no | yes | no |
| 1 | 480.8 | 480.8 | 480.8 |
| 2 | 480.8 | 461.0 | 480.8 |
| 3 | 425.4 | 425.4 | 425.4 |
| 4 | 416.5 | 416.5 | 416.5 |
| 5 | 410.7 | 410.7 | 410.9 |
| 6 | 394.6 | 394.6 | 394.6 |
| 7 | 394.6 | 388.9 | 388.9 |

Table 2: Disparities for galaxy data (unequal variances)

| $k$ | $-2\log L$ | | |
|---|---|---|---|
| Software | R | R | GLIM4 |
| Damping | no | yes | no |
| 1 | 480.8 | 480.8 | 480.8 |
| 2 | 440.7 | 440.7 | 440.7 |
| 3 | 407.0 | 407.0 | 407.0 |
| 4 | 395.4 | 395.4 | 395.4 |
| 5 | 380.9 | 380.9 | 392.3 |
| 6 | 365.2 | 365.2 | 365.2 |
| 7 | 363.0 | 359.9 | 363.0 |

(GLIM4 values in both tables from Aitkin, 2001)

Table 1 and 2 also provide the disparities as reported by Aitkin (2001), who used an equivalent NPML implementation in GLIM4. Comparing the first with the third column in Table 2, one notices that the disparity we obtained for the five mass-point model (380.9) is much better than the value 392.3 given by Aitkin. To investigate this, consider Figure 2 (middle, solid line), which shows the disparity $-2\ell$ as function of $d$. One realizes that the optimal solution is found if and only if $d$ is set to exactly $0.14$. This can certainly be overlooked easily, as apparently had happened. We re-checked this and fitted this model in GLIM4, and observed exactly the same behavior. However, the same value of $d$ does not necessarily and generally need to lead to the same disparity in GLIM4 and R. This is not too surprising, as small numerical differences, e.g. due to rounding, might heavily influence the results of a chaotic system, as the 'classical' EM algorithm seems to be. This may also explain why the 7-mass-point solution for equal variances with the disparity 388.9 reported by Aitkin could not be found at all by our R function using this 'classical' EM algorithm. We also re-checked this in GLIM4 and could not reproduce it either, though it certainly exists, as will be seen in the next section.

The properties of the EM trajectories seem to be not very fortunate. They suggest that selection of starting points or $d$ is rather a game of luck than something that can be optimized or steered. We try to improve this situation in the next section.

# 3 Damping the EM Algorithm

The crux of the matter is formula (7). Note that the standard deviation actually enters twice in $f_{ik}^{(0)}$, namely once by means of the extended linear predictor, and once as a distribution parameter. Consider the right middle picture in Figure 1. At iteration 0, each of these trajectories have to be imagined as the center of a normal distribution with mean $z_k^{(0)}$ and standard deviation $\sigma^{(0)}$. Obviously, when the mixture component means are very close, the $f_{ik}$ get blurred and lose their discriminatory power, as do the $w_{ik}$ as a consequence as well. This is in line with results from Ma and Xu (2005), who observed that the convergence properties of the EM algorithm deteriorate when the overlap between components of a Gaussian mixture increases. Thus, one has to adjust the standard deviations of the mixture components after extending the linear predictor. Fortunately, there

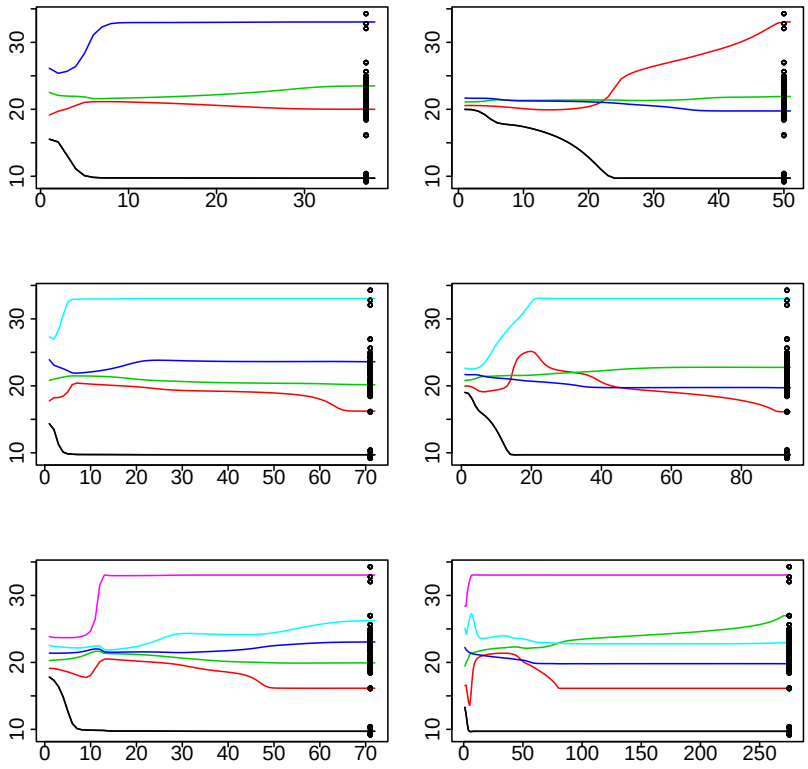

Figure 1: EM Trajectories for galaxy data with equal (left) and unequal (right) component standard deviations $\sigma_k$, $k = 1, \ldots, K$, $K = 4, 5, 6$ (from top to bottom). On the right hand side of each plot the velocities $y_i$ are depicted.

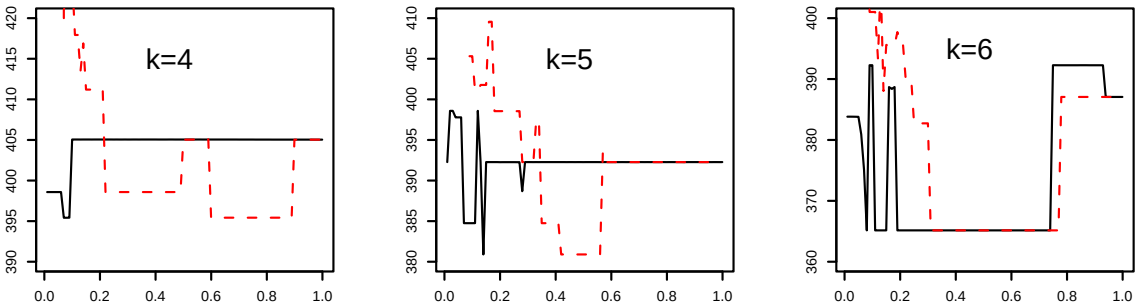


Figure 2: Disparity $-2\ell$ as a function of $d$ for $K = 4, 5$, and 6 with unequal variances (galaxy data). Solid line: undamped, dashed line: damped (see Section 3).

is an obvious way how to do that. Considering (5), one sees that the Gauss-Hermite mass points are actually multiplied by a 'working standard deviation' $\sigma_d^{(0)} = d \cdot \sigma^{(0)}$, which has consequentially to be used *at any occurrence in the likelihood*. Thus, the first point of improvement that we propose is to set

$$f_{ik}^{(0)} = f(y_i|h(\eta_i^{(0)} + \sigma_d^{(0)} g_k), \beta^{(0)}, (\sigma_d^{(0)})^2) = f(y_i|h(\eta_i^{(0)} + d\sigma^{(0)} g_k), \beta^{(0)}, d^2 \cdot \phi^{(0)}) \,. \quad (8)$$

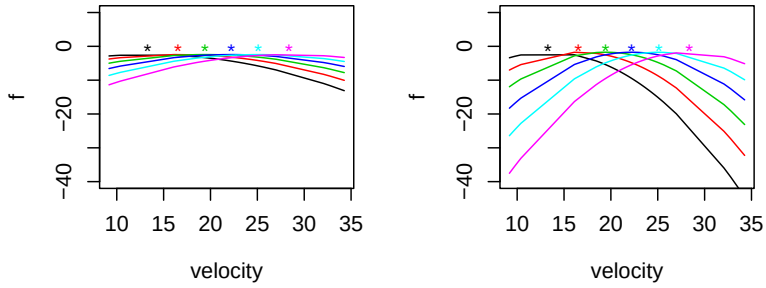


Figure 3: $f_{ik}^{(0)}$, $k = 1, \ldots, K$, $K = 6$, against velocity $y_i$ (galaxy data). Points (*) indicate starting points $z_k^{(0)}$. Left: Traditionally extended predictor without damping, right: damped.

Figure 3 illustrates the difference between the undamped and damped $f_{ik}^{(0)}$. In the further cycles damping has to be continued, as one otherwise just would have transported the problem from the first to the second cycle. As one has to get back to the 'true' likelihood at some point, the amount of 'damping' effected by the parameter $d$ has to be steadily reduced. This can be achieved by setting at iteration $j$

$$f_{ik}^{(j)} = f(y_i | h(x_i'\beta^{(j)} + z_k^{(j)}), \beta^{(j)}, d_j^2 \cdot \phi_{(k)}^{(j)}) \,, \tag{9}$$

with

$$d_j = 1 - (1-d)^{j+1}$$

and $\phi_{(k)}^{(j)} = \left(\sigma_{(k)}^{(j)}\right)^2$. Thus, one has $d_0 = d$ and $d_j \to 1$ $(j \to \infty)$ for $0 < d \leq 1$, implying that the likelihood converges to the 'true' (i.e. undamped) likelihood when the number of iterations gets large. To ensure a good approximation, the minimum number of iterations is set to ten in the current implementation. For instance for $d = 0.5$, one has already $d_{10} = 0.99951$ after 10 iterations.

We now apply the method described above to the galaxy data. We show the EM trajectories for a disparity-minimizing choice of $d$ in Figure 4. One observes that, for equal and unequal variances, the smoothness of the EM trajectories improves significantly. Fluctuations are reduced and the unintuitive crossings are avoided. As a by-product, the number of iterations necessary to achieve convergence falls significantly, for $k \geq 5$ dramatically. In some few cases (e.g. for $k = 2$ with equal variances) one even achieves superior solutions which could not been found without damping (see Table 1 and 2).

Moreover, a crucial advantage of the new method is that the sensitivity of the optimal solution to $d$ is reduced, as is obvious from Figure 2. Also the previously critical 7-mass point solution for equal variances, 388.9, is now achieved for a comfortable range of $d$ with width 0.22. It should be noted that we observed for the 'damped' EM algorithm a tendency to get trapped in so-called *likelihood spikes* (Aitkin et al., 2005, p. 428; Biernacki and Chretien, 2003) for unequal variances and very small values of $d$ (see Figure 2), as the damping procedure tends to hamper the EM trajectories from leaving a spike if they get caught in it in the first iterations. This can be avoided by setting a lower bound for the $\sigma_k$ (as by default in the `GLIM4` macro `normvar`), or by allowing a small amount of smoothing of the $\sigma_k$ among components, e.g. using the discrete kernel from Aitchison

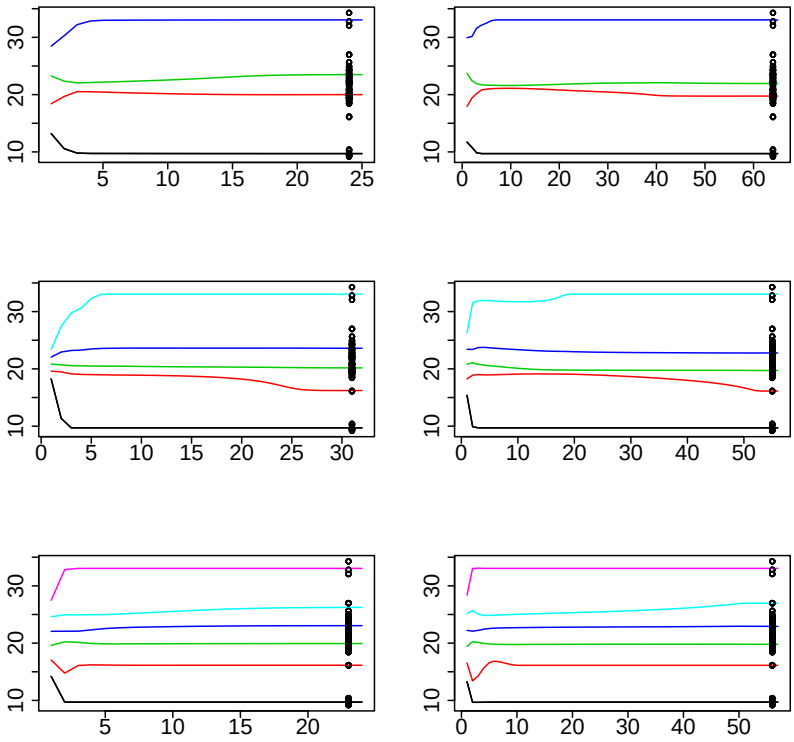


Figure 4: Damped EM trajectories for galaxy data with equal (left) and unequal (right) standard deviations $\sigma_k$, $k = 1, \ldots, K$, $K = 4, 5, 6$ (from top to bottom).

and Aitken (1976): $W(x, y|\lambda) = \lambda$ for $y = x$, and $(1 - \lambda)/(K - 1)$ otherwise. The latter possibility is implemented in the R function mentioned in the introduction. However, for a value of $d$ leading to a stable solution, which is nearly the entire range of $d$ as can be seen from Figure 2, neither of these measures is necessary.

# 4 Example: Gamma-Model for Hospital-Stay-Data

In this section we analyze a data set taken from Rosner (2000, p. 39) which is a sample from a larger data set collected on persons discharged from a Pennsylvania hospital as part of a retrospective chart review of antibiotic use in hospitals. A pre-analysis uncovered that only the covariates age and temp1 have a significant influence on the duration of hospital stay, where temp1 denotes the first measured temperature following admission, measured in Fahrenheit. The data set, reduced to these two covariates, is given in Table 3.

The distribution of the responses is highly skewed, as shown in Figure 5. Hence, a gamma distribution $\Gamma(\nu, \nu/\mu)$ with density $f(y|\mu, \nu) = (\nu/\mu)^\nu y^{\nu-1} e^{y\nu/\mu}/\Gamma(\nu)$, as is common for waiting time and duration problems, is more suitable than a normal model. We work with a logarithmic link function, i.e. $h^{-1}(\cdot) = \log(\cdot)$. Up to equation (6), NPML works as before. The damping procedure introduced for the normal model is now adapted straightforwardly from (8) and (9).

Since the dispersion parameter is now given by $\phi = 1/\nu$, the damped 'working shape

Table 3: Duration stay, age, and temperature at admission.

| ID | duration | age | temp1 | ID | duration | age | temp1 | ID | duration | age | temp1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 5 | 30 | 99.0 | 10 | 3 | 50 | 98.0 | 18 | 4 | 69 | 98.0 |
| 2 | 10 | 73 | 98.0 | 11 | 9 | 59 | 97.6 | 19 | 3 | 47 | 97.0 |
| 3 | 6 | 40 | 99.0 | 12 | 3 | 4 | 97.8 | 20 | 7 | 22 | 98.2 |
| 4 | 11 | 47 | 98.2 | 13 | 8 | 22 | 99.5 | 21 | 9 | 11 | 98.2 |
| 5 | 5 | 25 | 98.5 | 14 | 8 | 33 | 98.4 | 22 | 11 | 19 | 98.6 |
| 6 | 14 | 82 | 96.8 | 15 | 5 | 20 | 98.4 | 23 | 11 | 67 | 97.6 |
| 7 | 30 | 60 | 99.5 | 16 | 5 | 32 | 99.0 | 24 | 9 | 43 | 98.6 |
| 8 | 11 | 56 | 98.6 | 17 | 7 | 36 | 99.2 | 25 | 4 | 41 | 98.0 |
| 9 | 17 | 43 | 98.0 | | | | | | | | |

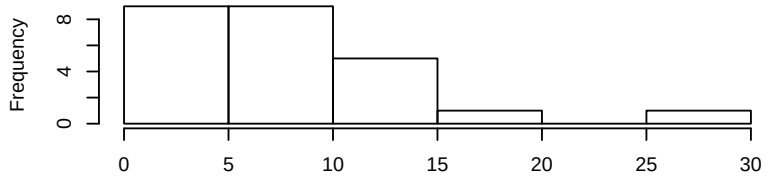


Figure 5: Histogram of duration

parameter' at iteration $j$, $j \geq 0$, is obtained by

$$\nu_d^{(j)} = \frac{1}{d_j^2}\,\nu^{(j)}\,,$$

whereby the shape parameter is obtained from the initial GLM (i.e. $\nu^{(0)} = 1/\phi^{(0)}$) in the initial cycle , and is estimated by

$$\nu^{(j)} = \frac{\sum_{i,k} w_{ik}^{(j)}}{\sum_{i,k} w_{ik}^{(j)} \left(\frac{y_i - \mu_{ik}^{(j)}}{\mu_{ik}^{(j)}}\right)^2} \tag{10}$$

with $\mu_{ik}^{(j)} = \exp(x_i'\beta^{(j)} + z_k^{(j)})$, in the subsequent cycles. In the case of unequal shape parameters $\nu_k$, the summations in (10) are taken for fixed $k$.

Firstly, a simple GLM with gamma-distributed response was fitted to the hospital data, using age and temp1 as covariates. The shape parameter was assumed to be constant over components. This model led to a deviance of 5.785, or correspondingly, to a disparity $-2\ell = 135.2$. Allowing for overdispersion improved this result drastically, yielding the disparity 121.3 for three mass points. Note that, without using the damping procedure, the disparity only fell to 134.5, as illustrated in Figure 6. Thus, the large overdispersion in the data would not have been captured by the undamped EM algorithm.

Given the large drop in disparity for overdispersion, one could ask whether the covariates are necessary at all. Fitting a GLM without covariates yields the disparity 145.3. The best overdispersed model is then a four mass point model with disparity 136.7 (damped) or 137.4 (undamped). However, the disparity 136.7 is still worse than the value 121.3

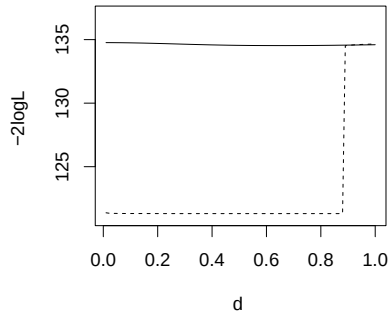


Figure 6: Dependence of disparity on $d$ for 3-mass-point model with covariates (solid: undamped, dotted: damped)

obtained in the presence of covariates, meaning that age and temp1 are still relevant when overdispersion is properly modelled.

By analogy to the unequal variances for the normal model, one can fit a model with unequal shape parameters. However, this turns out to be very unstable, as likelihood spikes occur here as well, with the shape parameter tending to infinity. Thus, smoothing across the component shape parameters with a discrete kernel here is not simply a minor improvement, but is essential to enable convergence (whether damping is applied or not). The best solution without covariates was achieved by a damped four mass point fit with smoothing parameter $\lambda = 0.9$, yielding $-2\ell = 135.3$ and shape parameters 14.5, 44.4, 30.8, 35.8. For comparison, the overall shape parameter for a four mass point model is 21.9, and for the null model 2.26.

# 5 Conclusion

We have set up some guidelines on how to perform NPML estimation for exponential families with unknown dispersion parameter and illustrated them by means of the normal and gamma model. There seems to be no reason why the introduced techniques should not be applied to other exponential family distributions, as e.g. Inverse Gaussian $IG(\mu, \lambda)$, where the dispersion parameter is given by $\phi = 1/\lambda$. The damping procedure would then be applied on $\lambda$ just as was outlined for the shape parameter of the gamma distribution.

**Acknowledgements**

This work was supported by Science Foundation Ireland Basic Research Grant 04/BR/M0051. The authors are grateful to M. Aitkin for providing the Galaxy data and the `GLIM4` macro `normvar`, to `R`. Darnell for providing his `R` copy of the `GLIM4` macro `alldist`, and to N. Sofroniou for providing his `GLIM4` macro `tolfind`.

# References

Abramowitz, M., and Stegun, I. A. (1970). *Handbook of Mathematical Functions*. Dover.

Aitchison, J., and Aitken, C. G. G. (1976). Multivariate binary discrimination by kernel method. *Biometrika*, *63*, 413-420.

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

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

Aitkin, M., and Francis, B. (1995). Fitting overdispersed generalized linear models by nonparametric maximum likelihood. *The GLIM Newsletter*, *25*, 37-45.

Aitkin, M., Francis, B., and Hinde, J. (2005). *Statistical Modelling in GLIM 4*. Oxford, UK: Oxford Statistical Science Series.

Biernacki, C., and Chretien, S. (2003). Degeneracy in the maximum likelihood estimation of univariate Gaussian mixtures with EM. *Statistics & Probability Letters*, *61*, 373-382.

Böhning, D. (1999). *Computer-Assisted Analysis of Mixtures and Applications. Meta-Analysis, Disease Mapping and others*. Boca Raton: Chapman & Hall / CRC.

Böhning, D., Schlattmann, P., and Lindsey, B. (1992). Computer-assisted analysis of mixtures (C.A.MAN): statistical algorithms. *Biometrics*, *48*, 283-303.

Hinde, J. (1982). Compound Poisson regression models. In R. Gilchrist (Ed.), *GLIM 82: Proceedings of the International Conference on Generalized Linear Models* (Vol. 14, p. 109-121). New York: Springer-Verlag.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, *73*, 805–811.

Ma, J., and Xu, L. (2005). Asymptotic convergence properties of the EM algorithm with respect to overlap in the mixture. *Neurocomputing*, *68*, 105-129.

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

Rosner, B. (2000). *Fundamentals of Biostatistics*. Duxbury, CA, USA: Duxbury Thomson Learning.

Authors' address:

Jochen Einbeck and John Hinde
Department of Mathematics
National University of Ireland, Galway
Ireland
E-mail: jochen.einbeck@nuigalway.ie
and john.hinde@nuigalway.ie