Statistics and Computing **13**: 257–265, 2003

The EM algorithm with gradient function update for discrete mixtures with known (fixed) number of components

DANKMAR BÖHNING

Free University Berlin/Humboldt University at Berlin, Joint Center for Health Sciences and Humanitites, Biometry and Epidemiology, Fabeckstr. 60-62, Haus 562, 14195 Berlin, Germany
boehning@zedat.fu-berlin.de

Received August 2002 and accepted January 2003

The paper is focussing on some recent developments in nonparametric mixture distributions. It discusses nonparametric maximum likelihood estimation of the mixing distribution and will emphasize gradient type results, especially in terms of global results and global convergence of algorithms such as vertex direction or vertex exchange method. However, the NPMLE (or the algorithms constructing it) provides also an estimate of the number of components of the mixing distribution which might be not desirable for theoretical reasons or might be not allowed from the physical interpretation of the mixture model. When the number of components is fixed in advance, the before mentioned algorithms can not be used and globally convergent algorithms do not exist up to now. Instead, the EM algorithm is often used to find maximum likelihood estimates. However, in this case multiple maxima are often occuring. An example from a meta-analyis of vitamin A and childhood mortality is used to illustrate the considerable, inferential importance of identifying the correct global likelihood. To improve the behavior of the EM algorithm we suggest a combination of gradient function steps and EM steps to achieve global convergence leading to the EM algorithm with gradient function update (EMGFU). This algorithms retains the number of components to be exactly k and typically converges to the global maximum. The behavior of the algorithm is highlighted at hand of several examples.

Keywords: mixture models, globally convergent algorithms, multiple maxima

1. The occurrence of mixtures

Mixture distributions occur in a very natural way. Suppose that for some random variate X of interest a probability density $f(x, \lambda)$ is valid, where λ is some real parameter. Suppose further that the population is *heterogeneous* in the sense that there exist, say, k subpopulations with parameter values $\lambda_1, \ldots, \lambda_k$. If sampling ignores the subpopulation membership then any of the k parameters could be valid and consequently the likelihood of observation x becomes

$$f(x) = p_1 f(x, \lambda_1) + \cdots + p_k f(x, \lambda_k) \tag{1}$$

where p_j represents the proportion of subpopulation j in the general population, $j = 1, \ldots, k$. The ignorance of population heterogeneity in terms of sampling is frequently unavoidable, since the covariate (representing the heterogeneity) is unknown or difficult to measure. We will denote the *mixing distribution* giving weight p_j to λ_j for $j = 1, \ldots, k$ by

$$P = \begin{pmatrix} \lambda_1 & \lambda_2 & \cdots & \lambda_k \\ p_1 & p_2 & \cdots & p_k \end{pmatrix} \tag{2}$$

and, consequently, indicate the dependency of $f(x)$ in (1) by $f(x) = f(x, P)$. $f(x, P)$ is the *mixture density* and $f(x, \lambda)$ the *mixture kernel*. Given a sample of size n, the log-likelihood is provided as

$$l(P) = \sum_{i=1}^{n} \log[f(x_i, P)] = \sum_{i=1}^{n} \log\left[\sum_{j=1}^{k} f(x_i, \lambda_j) p_j\right] \tag{3}$$

It is one of the important aspects of inference in mixture distributions to maximize the log-likelihood (3). Two cases have to be *clearly* distinguished. For one, the *number of mixture components* k might be fixed in advance,

Table 1. *Number of Death Notices x_i in the TIMES 1910–1912*

Count x_i	Frequency
0	162
1	267
2	271
3	185
4	111
5	61
6	27
7	8
8	3
9	1

and thus by, considered as known. For two, the number of components k might be itself unknown and part of the estimation process. In this case the so-called *non-parametric maximum-likelihood estimator* (NPMLE) may be considered. The name goes back to Laird (1978). Before we review some of the major algorithms we include a popular example for illustration (Hasselblad 1969, Titterington, Smith and Makov 1985).

1.1. *Example 1*

The following data are the number of death notices which appeared in the newspaper times between 1910 and 1912 for women aged 80 years and above (Table 1).

Frequently, for count data a Poisson distribution is used as mixture kernel:

$$f(x, \lambda) = Po(x, \lambda) = e^{-\lambda}\lambda^x/x! \tag{4}$$

Since a simple Poisson distribution provides a non-adequate fit, Hasselblad (1969) suggested a two-component mixture ($k = 2$), leading to the log-likelihood

$$l(P) = \sum_{i=1}^{n} \log\left[\sum_{j=1}^{k} Po(x_i, \lambda_j)p_j\right] \tag{5}$$

and the results of the maximum likelihood estimation (with $k = 2$ fixed) provides $\hat{P} = \begin{pmatrix} 1.2561 & 2.6634 \\ 0.3599 & 0.6401 \end{pmatrix}$. The two components are usually interpreted as different mortality patterns in winter and summer.

2. Global maximization

We consider the log-likelihood l in the *convex* set of all discrete probability distributions P. Note that this implies that the number of components k is *not* fixed in advance. This makes l to be *concave*.

2.1. *Gradient function*

As a major tool the directional derivative at P in the direction Q is used which is defined as:

$$\Phi(P, Q) = \lim_{\alpha \to 0} = [l((1-\alpha)P + \alpha Q) - l(P)]/\alpha \tag{6}$$

Note that (6) can be simply written as $\Phi(P, Q) = \sum_i f(x_i, Q)/f(x_i, P) - n$. The directional derivative becomes particularly simple for one-point probability mass directions $Q_\lambda : \Phi(P, Q_\lambda) = \sum_i f(x_i, \lambda)/f(x_i, P) - n$. This leads in a natural way to the gradient function as a normalized version of the directional derivative into the direction of the vertices of the probability simplex:

$$d(\lambda, P) = \frac{1}{n}\sum_{i=1}^{n} \frac{f(x_i, \lambda)}{f(x_i, P)} \tag{7}$$

In the example of the Poisson $f(x, \lambda) = Po(x, \lambda)$ the gradient function is simply $d(\lambda, P) = \frac{1}{n}\sum_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{\sum_j p_j e^{-\lambda_j}\lambda_j^{x_i}}$.

The *general mixture maximum likelihood theorem* (Lindsay 1983, Böhning 1982) can now easily be stated as: $\hat{P}$ is NPMLE if and only if $1 \geq d(\lambda, \hat{P})$ for all λ in the parameter space. In addition, $d(\lambda, \hat{P}) = 1$ for all mass points of $\hat{P}$ with non-zero mass.

This theorem is very useful in checking *candidates* for optimality. Consider the data set given in Table 2 which have been simulated from a single Poisson distribution with parameter 5.

It turns out that $\bar{x} = 4.78$ is the NPMLE, as it is easily verified that $1 \geq d(\lambda, \bar{x})$ for all λ. This implies that there is no need for further algorithmic iteration.

A second popular example is provided by the accident insurance data given in Thyrion (1960) and used by Simar (1976) in a pioneering paper on NPMLE for mixtures of Poisson distributions. The data are given in Table 3.

Simar (1976) provided $\hat{P}_{\text{Simar}} = \begin{pmatrix} 0.089 & 0.580 & 3.176 & 3.669 \\ 0.7600 & 0.2362 & 0.0037 & 0.0002 \end{pmatrix}$ as the NPMLE candidate which has been mentioned in the literature ever since, for example see Carlin and Louis (1996, p. 74, Table 3.2). However, it is easily verified that $\hat{P}_{\text{Simar}}$ is *not* NPMLE, since $d(x, \hat{P}_{\text{Simar}}) > 1$ for $x = 0$; see also Fig. 1. In fact, the NPMLE is provided in Leroux (1992) to be $\hat{P} = \begin{pmatrix} 0. & 0.3356 & 2.5454 \\ 0.4184 & 0.5730 & 0.0087 \end{pmatrix}$ (see also Böhning, 2000).

Table 2. *Simulated data set of size $n = 100$ from homogeneous Poisson distribution with $\lambda = 5$*

x_i	1	2	3	4	5	6	7	8	9	10
Frequency	2	10	17	20	19	12	10	4	4	2

Table 3. *Accident data of Thyrion (1960) used by Simar (1976)*

x_i	0	1	2	3	4	5	6	7
Frequency	7840	1317	239	42	14	4	4	1

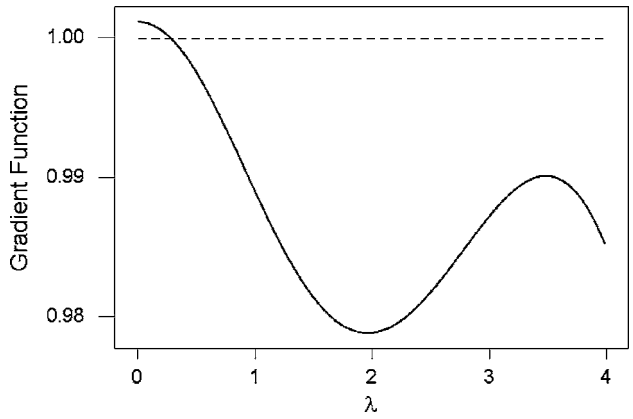

Fig. 1. *Gradient function $d(\lambda, P)$ for $P = P_{Simar}$ and mixture of Poissons for accident data of Table 3*

2.2. *Globally convergent algorithms*

Unfortunately, the general mixture maximum likelihood theorem is not helpful in *constructing* the NPMLE. Algorithms are required to accomplish this objective. One of the earliest is the *vertex direction method* (VDM). For any current discrete mass distribution P convex combinations $(1-\alpha)P + \alpha Q_\lambda$ are formed.

Vertex direction method (VDM)

- Step 0. Choose a discrete mass distribution $P = \binom{\lambda_1 \ \lambda_2 \ \cdots \ \lambda_k}{p_1 \ p_2 \ \cdots \ p_k}$ with arbitrary number of components k.
- Step 1. Determine $\lambda_{\max}$ to maximize $d(\lambda, P)$ in λ.
- Step 2. $\alpha_{\max}$ is found to maximize the log-likelihood $l((1-\alpha)P + \alpha Q_{\lambda_{\max}})$ in α, e.g. on the line connecting current P ($\alpha = 0$) with $Q_{\lambda_{\max}}$
- Step 3. Set $P = (1-\alpha)P + \alpha Q_{\lambda_{\max}}$ and go to Step 1.

In Step 1, $\lambda_{\max}$ is found to maximize the gradient function $d(\lambda, P)$ in λ. The intuition is to look for the vertex direction of steepest ascent. In Step 2, $\alpha_{\max}$ is found to maximize the log-likelihood $l((1-\alpha)P + \alpha Q_{\lambda_{\max}})$ in α, e.g. on the line connecting current P ($\alpha = 0$) with $Q_{\lambda_{\max}}$ ($\alpha = 1$). Then, current P is set to be $(1-\alpha_{\max})P + \alpha_{\max} Q_{\lambda_{\max}}$ and Steps 1 and 2 are repeated until convergence. To illustrate, consider $\binom{\lambda_1 \ \lambda_2 \ \lambda_3}{p_1 \ p_2 \ p_3}$ and let $\lambda_{\max} = \lambda_4$. Then, $(1-\alpha_{\max})P + \alpha_{\max} Q_{\lambda_{\max}} = \binom{\lambda_1 \ \lambda_2 \ \lambda_3 \ \lambda_4}{p_1 \ p_2 \ p_3 \ \alpha_{\max}}$, with $p_j' = (1-\alpha_{\max})p_j$ for $j = 1, 2, 3$. The disadvantages of the VDM are that it is very slow in convergence and has the tendency to generate clusters of components. A largely improved alternative is the *vertex exchange method* (VEM). In Step 1 of the VEM, not only $\lambda_{\max}$ is found to maximize the gradient function $d(\lambda, P)$ in λ, but also $\lambda_{\min}$ is found to minimize the gradient function $d(\lambda, P)$ under those λ_j of the current P which receive positive weight.

Vertex exchange method (VEM)

- Step 0. Choose a discrete mass distribution $P = \binom{\lambda_1 \ \lambda_2 \ \cdots \ \lambda_k}{p_1 \ p_2 \ \cdots \ p_k}$ with arbitrary number of components k.
- Step 1. Determine $\lambda_{\max}$ to maximize $d(\lambda, P)$ in λ, where λ is varying in the whole parameter space, and $\lambda_{\min}$ such that $d(\lambda_{\min}, P)$ is smallest in the set with k elements

$$\{d(\lambda_1, P), d(\lambda_2, P), \ldots, d(\lambda_k, P)\},$$

  where $\lambda_1, \ldots, \lambda_k$ receive positive weight in P.
- Step 2. $\alpha_{\max}$ is found to maximize the log-likelihood $l(P + \alpha P(\lambda_{\min})[Q_{\lambda_{\max}} - Q_{\lambda_{\min}}])$ in α.
- Step 3. Set $P = P + \alpha_{\max} P(\lambda_{\min})[Q_{\lambda_{\max}} - Q_{\lambda_{\min}}]$ and go to Step 1.

The intuition is to look not only for vertex directions of steepest ascent, but also for those mass points which are "bad" support points with respect to the gradient function. The VEM then forms $P + \alpha P(\lambda_{\min})[Q_{\lambda_{\max}} - Q_{\lambda_{\min}}]$ where α is again a line maximizer of the likelihood on the line connecting P with $P + P(\lambda_{\min})[Q_{\lambda_{\max}} - Q_{\lambda_{\min}}]$. Note that if $\alpha = 1$ the "bad" support point $\lambda_{\min}$ is *exchanged* with $\lambda_{\max}$; this motivated the name vertex exchange method. For details see Böhning (2000). As an illustration, consider again $\binom{\lambda_1 \ \lambda_2 \ \lambda_3}{p_1 \ p_2 \ p_3}$ and let $\lambda_{\max} = \lambda_4$ and $\lambda_{\min} = \lambda_2$, the latter in the current support of P. Then, we have $P + \alpha_{\max} P(\lambda_{\min})[Q_{\lambda_{\max}} - Q_{\lambda_{\min}}] = \binom{\lambda_1 \ \lambda_2 \ \lambda_3 \ \lambda_4}{p_1 \ (1-\alpha_{\max})p_2 \ p_3 \ \alpha_{\max} p_2}$. Note that if $\alpha_{\max} = 1$, then $P + P(\lambda_{\min})[Q_{\lambda_{\max}} - Q_{\lambda_{\min}}] = \binom{\lambda_1 \ \lambda_3 \ \lambda_4}{p_1 \ p_3 \ p_2}$, and the component λ_2 is *exchanged* with the new component λ_4. The VEM is converging much better than the VDM and it is able to discard "bad" components rather easily. Both, VDM and VEM provide globally convergent algorithms in the sense that they converge from any initial disrete probability distribution to the NPMLE.

3. Number of components fixed: Globally convergent algorithms

The algorithms discussed in the previous section will deliver some estimate $\hat{k}$ of the number of components k. Frequently however, it is desired to keep the number of components k constant. The reasons for doing so might be manifold, a few of them are mentioned as follows.

3.1. *Selection procedures for the number of components*

The statistical procedure might require to fix the number of components. Even so an estimate $\hat{k}$ of k has been found, this estimate might be unnecessarily large, that is a smaller value of k might lead to a similar likelihood. Thus, besides $\hat{k}$ values like $\hat{k} - 1$, $\hat{k} - 2$, ... are of interest and compared with respect to their log-likelihood or BIC - value (see also Leroux 1992).

We would like to illustrate the inferential consequences using the *wrong* maximum likelihood by means of an example from meta-analyis. Fawzi *et al.* (1993) study the effect of Vitamin A supplementation and childhood mortality in preschool children. We reproduce their Table 4 as our Table 4. All studies are community-randomized trials from South-Asia or Sout-East Asia, besides the second study which is from Northern Sudan.

Table 4. *Mortality in community-based trials of Vitamin A supplementation in children aged 6 to 72 months*

Location	Obs.-Time	Vitamin A[a]	Control[a]	log-RR	Variance
Sarlahi (Nepal)	12	152; 14487	210; 14143	−0.34726	0.011341
Northern Sudan	18	123; 14446	117; 14294	0.03943	0.016677
Tamil Nadu (India)	12	37; 7764	80; 7655	−0.78525	0.039527
Aceh (Indonesia)	12	101; 12991	130; 12209	−0.31450	0.017593
Hyderabad (India)	12	39; 7691	41; 8084	−0.00017	0.050031
Jumla (Nepal)	5	138; 3786	167; 3411	−0.29504	0.013234
Java (Indonesia)	12	186; 5775	250; 5445	−0.35455	0.009376
Bombay (India)	42	7; 1784	32; 1644	−1.60155	0.174107

[a]Entries are number of child deaths and number of children under risk.

The *incidence density* was estimated according to $\widehat{\mathrm{ID}} = E/T$, where E are the number of child deaths and T is the person time, caluclated as the product of *children under risk* and *years of observation*. Consequently, the rate ratio was estimated as $\widehat{\mathrm{RR}} = \frac{\widehat{\mathrm{ID}}_A}{\widehat{\mathrm{ID}}_C}$, where the index refers to intervention with Vitamin A supplementation (A) and control (C). For the variance of the log-rate-ratio the large sample formula $\mathrm{Var}(\widehat{\log \mathrm{RR}}) = 1/E_A + 1/E_C$ is used, where the indices are as previously defined. To put the results in a nutshell, Vitamin A supplementation turns out to be beneficial, though the effect is more beneficial in some studies than in others. We observe effect heterogeneity which can be modelled using a mixture approach. Following Laird (1978) we model the effect measure $x_i = \log \widehat{\mathrm{RR}}_i$ for the i-th study as a mixture of normal densities $f(x_i) = p_1 f(x_i, \lambda_1) + \cdots + p_k f(x_i, \lambda_k)$, where $f(x_i, \lambda) = N(x_i, \lambda, \sigma_i^2)$ is the normal density with mean λ and variance σ_i^2 equal to the observed variance as provided in the last column of Table 4.

In Table 5 the log-likelihoods for mixture models with various number of components k are provided. For $k = 2$, depending on the initial value of the EM algorithm *considerable different* log-likelihoods are delivered. In addition, not only the log-likelihoods are affected, but also other criteria involving the likelihood such as the *Bayesian Information Criterion* defined as $\mathrm{BIC} = 2l(P^\infty) - (2k - 1)\log(n)$ which is frequently recommended as a guideline for selecting the number of components (McLachlan and Peel 2000). According to this guideline we would choose the model with the largest BIC-value. Set 1 given in column 2 of Table 5 (starting with equal weights on −1.6 and 0) delivers the correct maximum likelihood leading to a choice of $k = 2$ in the model, whereas the other two sets (set 2 gives equal weight to −0.5 and 0, set 3 equal weight to −1.6 and −0.5) would provide only local solutions leading to a choice of $k = 1$ for the number of components implying homogeneity. These inferential and, as a further result, substantial consequences in terms of the interpretation of the meta-analysis highlight the importance of identifying the correct maximum likelihood.

Table 5. *Log-likelihoods and BIC-values for different values of k in the meta-analysis of Fawzi et al. (1993)*

k	Set of initial values	$l(P^\infty)$[a]	BIC	No. of parameters
1	–	−5.00399	−12.0874	1
2	1	−2.73066	−11.6996	3
2	2	−3.23697	−12.7123	3
2	3	−3.10309	−12.4445	3
3	–	−1.56781	−13.5328	5
4	(NPMLE)	−1.19598	−16.9481	7

[a]∞ Indicates the parameter values at termination of EM algorithm.

3.2. *Model requirement for a fixed value of the number of components*

Sometimes the physiological or biological model requires certain values for k such as $k = 2$. Specifically, in population related (non-clinican) medical disciplines like public health mixture models are of high interest since they account for potential heterogeneity in large population studies. For example, it it well-known that the diabetis mellitus indicator BLOOD GLUCOSE experiences a typical two-component normal mixture when studied in large residential populations (in contrast to clinical populations). For details, see the work of Lim *et al.* (2001).

In the case of fixed number of components no globally convergent algorithm exists up-to-date.

Usually, it is recommended to use the EM algorithm (Dempster, Laird and Rubin 1977) with a number of different trial values. Firstly, we look at the EM algorithm for the mixture setting.

3.3. *EM algorithm for mixtures*

In the mixture setting, the complete-data likelihood is

$$\prod_{i=1}^{n}\prod_{j=1}^{k} f(x_i, \lambda_j)^{z_{ij}} p_j^{z_{ij}} \tag{8}$$

where z_{ij} are n unobserved realizations of k component-indicators $Z_{i1}, \ldots, Z_{ik}$. This leads to the complete-data

log-likelihood

$$l_{\text{com}}(P) = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\log(f(x_i,\lambda_j)) + \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\log(p_j) \quad (9)$$

and the well-known *E-Step*: $E(Z_{ij} \mid P, \mathbf{x}) = e_{ij} = \frac{f(x_i,\lambda_j)p_j}{\sum_{j'} f(x_i,\lambda_{j'})p_{j'}}$ and *M-Step* in which the z_{ij} in (9) are replaced by the their expected values e_{ij}. Maximizing this expected log-likelihood leads to new estimates as such: $p_j^{\text{new}} = \sum_i e_{ij}/n$ for the weights and λ_j^{new} where the new values for the λ_j-parameters will depend on the form of the density $f(x,\lambda)$. If for the single component model the MLE is the mean, then the form of new estimates for λ_j can be provided as: $\lambda_j^{\text{new}} = \sum_i e_{ij}x_i / \sum_i e_{ij}$.

3.4. *Multiple maxima*

The problem of occurrence of several local maxima is well-known for the setting described in the previous section, though it is seldom investigated in detail. Seidel, Mosler and Alker (2000) point out that the simulated null-distribution of the likelihood-ratio test depends on the choice of the initial value for the EM algorithm. As an illustration consider a sample of size 100 from a single component exponential distribution. For the data set itself see the appendix. The number of components is fixed to be $k = 2$. As Table 6 illustrates, depending on the initial value for the EM algorithm *considerable different* log-likelihoods are delivered. In addition, not only the log-likelihoods are affected, but also other criteria involving the likelihood such as the *Bayesian Information Criterion* (defined previously in Section 3.1) or *Akaike Information Criterion* defined as $AIC = 2l(P^\infty) - 2(2k-1)$ which are frequently chosen as a guideline for selecting the number of components (Leroux 1992, Celeux 2001, McLachlan and Peel 2000).

Table 7 incorporates the BIC-values as well and a variety of different BIC-values are obtained depending on the initial value of the EM algorithm.

Table 6. *Different likelihoods at convergence for different initial values (k = 2)*

Set	$\lambda_j^{(0)}$	$p_j^{(0)}$	λ_j^∞	p_j^∞	$l(P^\infty)$[a]
1	1.0000	0.5000	0.7296	0.5749	−73.3814
	2.0000	0.5000	0.8154	0.4251	
2	0.5000	0.5000	0.7590	0.4781	−73.3555
	1.0000	0.5000	0.7726	0.5219	
Extreme values					
3	0.0010	0.5000	0.0019	0.0235	−71.0982
	3.7000	0.5000	0.7845	0.9765	
Quartiles					
4	0.1800	0.5000	0.0239	0.0939	−69.0262
	1.2800	0.5000	0.8430	0.9061	
5	0.5000	0.5000	0.7552	0.5091	−73.3566
	1.5000	0.5000	0.7774	0.4909	

[a] ∞ Indicates the parameter values at termination of EM algorithm.

Table 7. *Log-likelihoods and BIC-values for different values of k*

k	Set of initial values[a]	$l(P^\infty)$	BIC	No. of parameters
1	$(\bar{x})$	−73.3549	−151.315	1
2	1	−73.3814	−160.578	3
2	2	−73.3555	−160.527	3
2	3	−71.0982	−156.012	3
2	4	−69.0262	−151.868	3
2	5	−73.3566	−160.529	3
3	(NPMLE)	−68.8691	−160.764	5

[a] According to Table 6.

3.5. *A globally convergent algorithm*

Typically, the EM algorithm is employed *not* using knowledge from the existing global maximization theory. The global optimization theory was reviewed in Section 2 in detail to lay the ground for some simple, theory-guided adjustments of the EM algorithm to circumvent local maxima. The idea is simply to combine the algorithmic approaches in Section 2 using the gradient function with the EM algorithm.

EM Algorithm with gradient function update (EMGFU)

- Step 0. Choose and fix the number of components k; choose arbitrary starting value $P = \left(\begin{smallmatrix}\lambda_1 & \lambda_2 & \cdots & \lambda_k \\ p_1 & p_2 & \cdots & p_k\end{smallmatrix}\right)$ for EM algorithm.
- Step 1. Use EM algorithm to provide at convergence $P_{\text{EM}} = P^\infty$.
- Step 2. Determine $\lambda_{\max}$ to maximize $d(\lambda, P_{\text{EM}})$ in λ.
- Step 3. Determine $\lambda_{\min}$ such that $l(P_{\text{EM}} + P_{\text{EM}}(\lambda_{\min})[Q_{\lambda_{\max}} - Q_{\lambda_{\min}}])$ is largest in the set with k elements

$$\{l(P_{\text{EM}} + P_{\text{EM}}(\lambda_j)[Q_{\lambda_{\max}} - Q_{\lambda_j}]) \mid j = 1,\ldots,k\},$$

  where $\lambda_1, \cdots, \lambda_k$ receive positive weight in P_{EM}.
- *Comment.* Note that in Step 3 exactly k values of the log-likelihood are computed. The point $\lambda_{\max}$ found in Step 2 is exchanged with each of the k component parameters λ_j, where j runs from 1 to k, and the associated log-likelihood is formed. Furthermore, note that in Step 3 the gradient function is not a suitable selection criterion, since for all k values of λ the gradient function coincides (see Theorem 2).
- Step 4. Let $P = P_{\text{EM}} + P_{\text{EM}}(\lambda_{\min})[Q_{\lambda_{\max}} - Q_{\lambda_{\min}}]$ (Exchange $\lambda_{\max}$ with $\lambda_{\min}$). If $l(P) > l(P_{\text{EM}})$, go to Step 1; otherwise stop.

Note that by forcing $\alpha = 1$ in Step 3, the number of components is always exactly k. Of course, by construction, there is guarantee of *monotonicity* and, thus by, convergence.

Theorem 1. *Any sequence of log-likelihoods created by the EMGFU converges* monotonically *to a stationarity point.*

In Step 3 of the EMGFU, the gradient function was *not* used as selection criterion, as it was suggested in the Vertex-Exchange-Method in Section 2. The reason for using the log-likelihood instead lies in the fact that the gradient function coincides for all λ-values obtained at termination of the EM-algorithm.

Theorem 2. *Let P_{EM} be the discrete probability distribution at convergence of the EM algorithm, λ_j and p_j the associated paramters for $j = 1, \ldots, k$. Then,*

$$d(\lambda_1, P_{\mathrm{EM}}) = d(\lambda_2, P_{\mathrm{EM}}) = \cdots = d(\lambda_k, P_{\mathrm{EM}}).$$

Proof: The *E-Step* is given as $E(Z_{ij} \mid P, \mathbf{x}) = e_{ij} = \frac{f(x_i,\lambda_j)p_j}{\sum_{j'} f(x_i,\lambda_{j'})p_{j'}}$, and from here, the *M-Step* $p_j^{\mathrm{new}} = \sum_i e_{ij}/n$. In particular, for $P = P_{\mathrm{EM}}$,

$$p_j^{\mathrm{new}} = \frac{1}{n}\sum_i \frac{f(x_i,\lambda_j)p_j}{\sum_{j'} f(x_i,\lambda_{j'})p_{j'}} = p_j$$

for all j from 1 to k. Dividing both sides by p_j gives

$$d(\lambda_j, P_{\mathrm{EM}}) = 1$$

for all j from 1 to k, which is statement of the theorem. □

Theorem 1 provides convergence, though there is no guarantee of convergence to a *global* maximum. However, we will demonstrate on empirical grounds that this simple adjustment of the EM algorithm by means of the gradient function provides a considerable improvement. In fact, no case has been observed where it failed to provide the global maximum for the fixed component case.

3.6. *Empirical evidence*

In this section we demonstrate how the *EM algorithm with gradient function update* works in practice. We consider the mixtures of exponentials with $k = 2$ components for the data set in Section 3.2. The results are provided in Table 8.

We start the EM algorithm with some initial values that had lead to some local solution with rather inferior likelihood (see Table 6). The gradient function is maximized near 0, at $\lambda_{\max} = 0.0020$. There are only $k = 2$ mass points and the one with the smaller gradient function is $\lambda_{\min} = 0.7590$. The latter is exchanged with 0.0020, and the EM algorithm started again. We get a highly improved likelihood. Repeating the process does not improve the likelihood anymore, therefore the algorithm terminates. The algorithm has been started from any of the sets used in Table 6, leading to log-likelihoods, identical to −69.0263. Though the EMGFU algorithm is a clear improvement of the EM algorithm for mixtures, there is no guaranteed proof of convergence to a global maximum. However, further empirical studies support the global convergence character of the EMGFU algorithm.

Table 8. *Illustration of EMGFU for mixtures of $k = 2$ exponentials*

Iteration	$\lambda_j^{(0)}$	$p_j^{(0)}$	λ_j^{∞}	p_j^{∞}	$l(P^{\infty})$[a]
1	0.5000	0.5000	0.7590	0.4781	−73.3555
	1.0000	0.5000	0.7726	0.5219	
$\lambda_{\max} = 0.0020$					
$\lambda_{\min} = 0.7590$					
2	0.0020	0.4781	0.0239	0.0939	−69.0263
	0.7726	0.5219	0.8430	0.9061	
$\lambda_{\max} = 0.0020$					
$\lambda_{\min} = 0.0239$					
3	No improvement in step 4 → stop!				−69.0263

[a]∞ Indicates the parameter values at termination of EM algorithm.

Table 9. *Different likelihoods at convergence for different initial values ($k = 3$)*

Set	$\lambda_j^{(0)}$	$p_j^{(0)}$	λ_j^{∞}	p_j^{∞}	$l(P^{\infty})$[a]
1	1.0000	0.3333	0.7620	0.4046	−73.3550
	2.0000	0.3333	0.7685	0.2944	
	3.0000	0.3333	0.7693	0.3044	
Extreme values					
2	0.0010	0.3333	0.0019	0.0235	−71.0983
	0.5700	0.3333	0.7831	0.5724	
	3.7000	0.3333	0.7863	0.4042	
Quartiles					
3	0.1800	0.3333	0.0239	0.0939	−69.0262
	0.5700	0.3333	0.8430	0.3740	
	1.2800	0.3333	0.8430	0.5321	

[a]∞ Indicates the parameter values at termination of EM algorithm.

Let us continue the discussion by considering $k = 3$ components for the mixture of exponentials.

Table 9 illustrates again that local maxima can easily occur. However, we will use this example to point out a further problem which we call the *dimension reduction* problem. Suppose we use set 1 to start the EMGFU. The results are provided in Table 10. Evidently, in Step 2 the two points 0.7685 and 0.7693 are collapsed by the EM-algorithm to *one* point, leading now to a mixture of two components only whereas we are interested in finding the maximum likelihood estimate in $k = 3$ components.

The solution of this reduction problem is provided by a *dimension adjustment step* in the EMGFU algorithm.

3.7. *Dimension adjustment*

We include a dimension adjustment in the EMGFU algorithm.

EM algorithm with gradient function update (EMGFU) and dimension adjustment

- Step 0, Step 2, Step 3, and Step 4 are as in the EMGFU of Section 3.5.
- Step 1. Use EM algorithm to provide at convergence $P_{\mathrm{EM}} = P^{\infty}$.

  Step 1.1. If the number of components $= k$, go to Step 2. *Dimension adjustment*:

Table 10. *Illustration of EMGFU for mixtures of $k = 3$ exponentials*

Iteration	$\lambda_j^{(0)}$	$p_j^{(0)}$	λ_j^{∞}	p_j^{∞}	$l(P^{\infty})$[a]
1	1.0000	0.3333	0.7620	0.4046	−73.3550
	2.0000	0.3333	0.7685	0.2944	
	3.0000	0.3333	0.7693	0.3044	
$\lambda_{\max} = 0.0020$					
$\lambda_{\min} = 0.7620$					
2	0.0020	0.4046	0.0239	0.0939	−69.0263
	0.7685	0.2944	0.8430	0.9061	
	0.7693	0.3044	–	–	
	$k = 3$ is reduced to $k = 2$				

[a] ∞ Indicates the parameter values at termination of EM algorithm.

Step 1.2. If the number of components $= k - 1$, determine $\lambda_{\max}$ to maxmimize $d(\lambda, P_{\mathrm{EM}})$ in λ and set $P = (1 - \alpha_{\max})P_{\mathrm{EM}} + \alpha_{\max} Q_{\lambda_{\max}}$ and go to Step 1. Here $\alpha_{\max}$ is chosen as in the vertex direction method (see Section 2.2). If $\alpha_{\max} > 0$ with $l(P) > l(P_{\mathrm{EM}})$ does *not* exist, then P_{EM} must be the NPMLE, and iteration stops.

The EMGFU is again monotonic and, consequently, the sequence of associated log-likelihoods has to converge. There is, however, no guarantee of global convergence.

We now apply this technique to the situation of Table 10 where the dimension reduction problem had occurred. The results are provided in Table 11.

In Step 3 of Table 11 the nonparametric maximum likelihood estimator

$$P_{\mathrm{NPMLE}} = \begin{pmatrix} 0.0017 & 0.0271 & 0.8419 \\ 0.0102 & 0.0825 & 0.9073 \end{pmatrix} \qquad (10)$$

has been reached since $1 \geq d(\lambda, P)$ for all λ as also can be seen in Fig. 2.

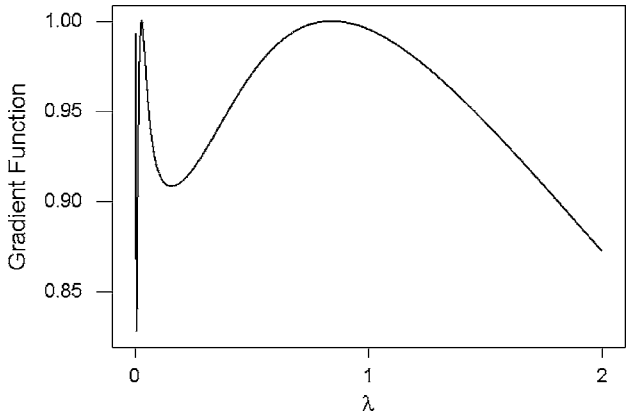

Fig. 2. *Gradient function $d(\lambda, P)$ for $P = P_{NPMLE}$ given in (10)*

In this case it is evident that the global maximum has been achieved and the algorithm is terminated. In other cases, the algorithm will be terminated if there is no further improvement in the likelihood.

Table 11. *Illustration of EMGFU with dimension adjustment for mixtures of $k = 3$ exponentials*

Iteration	$\lambda_j^{(0)}$	$p_j^{(0)}$	λ_j^{∞}	p_j^{∞}	$l(P^{\infty})$[a]
1	1.0000	0.3333	0.7620	0.4046	−73.3550
	2.0000	0.3333	0.7685	0.2944	
	3.0000	0.3333	0.7693	0.3044	
$\lambda_{\max} = 0.0020$	$\lambda_{\min} = 0.7620$				
(Vertex exchange step)					
2	0.0020	0.4046	0.0239	0.0939	−69.0263
	0.7685	0.2944	0.8430	0.9061	
	0.7693	0.3044	–	–	
$k = 3$ is reduced to $k = 2$					
$\lambda_{\max} = 0.0020$	$\alpha_{\max} = 0.0051$				
(Vertex direction step)					
3	0.0239	0.0934	0.0271	0.0825	−68.8691
	0.8430	0.9061	0.8419	0.9073	
	0.0020	0.0051	0.0017	0.0102	
	$d(\lambda_{\max}, P) = 1 \rightarrow$ stop				

[a] ∞ Indicates the parameter values at termination of EM algorithm.

In the dimension adjustment step (Step 1.2) it was assumed that the dimension reduction of the EM algorithm is from k to $k-1$. In full generality, it might be that several points are collapsed in which case the reduction is from k to $k-m$, where m is an integer with $0 < m < k$. Note that in this case, instead of *one* vertex direction step m vertex direction steps will be required.

3.8. *A suggestion for the choice of α_{max} in the dimension adjustment step*

Finally, we want to point out a special property of the log-likelihood which suggests a choice for $\alpha_{\max}$ which can be used in Step 1.2 (Diemension Adjustment of the EMGFU). Consider $\varphi(\alpha) = l((1-\alpha)P + \alpha Q_\lambda) = \sum_x \log[(1-\alpha)f(x, P) + \alpha f(x, \lambda)]$. The derivatives of φ do have a simple structure:

$$\varphi'(\alpha) = \sum_x \frac{f(x,\lambda) - f(x,P)}{(1-\alpha)f(x,P) + \alpha f(x,\lambda)}\bigg|_{\alpha=0} = \sum_x g(x,\lambda,P) \tag{11}$$

and

$$\begin{aligned}\varphi''(\alpha) &= -\sum_x \frac{[f(x,\lambda) - f(x,P)]^2}{[(1-\alpha)f(x,P) + \alpha f(x,\lambda)]^2}\bigg|_{\alpha=0} \\ &= -\sum_x g(x,\lambda,P)^2 \end{aligned} \tag{12}$$

where $g(x,\lambda,P) = \frac{f(x,\lambda)-f(x,P)}{f(x,P)}$. Note that $0 \geq (12)$ for all λ and x. The (one-step) Newton-Raphson correction is provided by

$$\alpha_{\max} \approx \frac{\sum_x g(x,\lambda,P)}{\sum_x g(x,\lambda,P)^2} \tag{13}$$

which is neccessarily positive if λ is a vertex of ascent, e.g. if $d(\lambda, P) > 1$.

Acknowledgments

I am grateful to Prof. Dr. Wilfried Seidel (University of the Armed Forces of Germany at Hamburg) for discussing and pointing out several critical issues which have lead to considerable improvements of the paper. I would also like to express my gratitude to Prof. Dr. Marco Alfò (Università "La Sapienza", Rome) for several discussions on the subject. Thanks go also to the Editor of *Statistics and Computing* for helpful comments and suggestions. This research is under current support of the *German Research Foundation*.

References

Böhning D. 1982. Convergence of Simar's algorithm for finding the maximum likelihood estimate of a compound Poisson process. Annals of Statistics 10: 1006–1008.

Böhning D. 2000. Computer-Assisted Analysis of Mixtures and Applications. Meta-Analysis, Disease Mapping and Others. Chapman & Hall/CRC, Boca Raton.

Carlin B.P. and Louis T.A. 2000. Bayes and Empirical Bayes Methods for Data Analysis. Chapman & Hall, London.

Celeux G. 2001. Different points of view for choosing the number of components in a mixture model. In: Govaert G., Janssen J., and Limnios N. (Eds.), Proceedings of the 10th International Symposium on Applied Stochastic Models and Data Analysis, June 12–15. Compiegne, pp. 21–28.

Dempster A.P., Laird N.M., and Rubin D.B. 1977. Maximum likelihood estimation from incomplete data via the EM algorithm (with Discussion). Journal of the Royal Statistical Society B 39: 1–38.

Fawzi W.W., Chalmers T.C., Herrera M.G., and Mosteller F. 1993. Vitamin A supplementation and child mortality. A meta-analysis. Journal of the American Medical Association 269: 898–903.

Hasselblad V. 1969. Estimation of finite mixtures of distributions from the exponential family. Journal of the American Statistical Association 64: 1459–1471.

Appendix: Data set used for testing the EMGFU algorithm

0.03302	0.67841	0.83678	1.70085	3.73222	0.39648	0.16839	0.57422	1.29458
0.12759	0.44600	0.68039	1.31812	0.47316	1.67348	0.22541	0.81522	0.54392
1.64572	0.81737	0.23003	1.47947	0.18865	0.56448	0.33223	1.14901	0.16381
1.80573	0.66226	1.30628	0.15858	0.05621	1.66521	1.01774	1.75035	0.62135
1.60808	0.76876	0.02127	0.92949	0.14542	0.26653	0.01962	0.04570	0.19571
0.18483	1.15779	1.27279	0.00297	0.96688	0.78516	0.51107	0.11811	1.83021
3.07632	0.28069	1.01281	0.34646	0.03557	0.65484	1.57239	0.02906	0.51749
0.06384	1.44599	1.01078	0.76734	0.05908	0.57213	2.34580	0.01476	1.38737
0.82217	0.01586	0.05073	0.27409	0.01410	1.33783	0.53023	0.38914	0.02472
0.32186	0.00151	1.84842	0.77284	2.26805	1.38125	0.56990	0.77199	0.42500
1.84390	0.25340	0.25842	1.54009	0.00125	1.70587	0.05284	1.10530	0.25739
0.41535								

These 100 data have been sampled from an exponential with parameter 1 and were used in Section 3 as test data for the algorithm EMGFU.

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

Leroux B.G. 1992. Consistent estimation of a mixing distribution. Annals of Statistics 20: 1350–1360.

Lim, T.-O., Bakri R., Morad Z., and Hamid M.A. 2001. Bimodality in blood glucose distribution: Is it universal? Preprint of the Clinical Research Centre, c/o Department of Nephrology, Kualalumpur Hospital, Kualalumpur, Malaysia.

Lindsay B.G. 1983. The geometry of mixture likelihoods, Part I: A general theory. Annals of Statistics 11: 783–792.

Mclachlan G. and Peel D. 2000. Finite Mixture Models. Wiley, New York.

Seidel W., Mosler K., and Alker M. 2000. A cautionary note on likelihood ratio tests in mixture models. Annals of the Institute of Statistical Mathematics 52: 481–487.

Simar L. 1976. Maximum likelihood estimation of a compound Poisson process. Annals of Statistics 4: 1200–1209.

Thyrion P. 1960. Contribution à l'étude du bonus pour non sinsitre en assurance automobile. Astin Bulletin 1: 142–162.

Titterington D.M., Smith A.F.M., and Makov U.E. 1985. Statistical Analysis of Finite Mixture Distributions. Wiley, New York.