The EM Algorithm with Gradient Function Update for Discrete Mixtures

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

The paper is focussing on some recent developments in nonparametric mixture distributions. When the number of components is fixed in advance 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).

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 $\lambda_1, ..., \lambda_k$. If sampling ignores the subpopulation membership then any of the k parameters could be valid and consequently the likelihood of observation x becomes

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

where p_j represents the proportion of subpopulation j in the general population, $j = 1, ..., k$. 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 $j = 1, ..., k$ by $P = \begin{pmatrix} \lambda_1 & \lambda_2 & \cdots & \lambda_k \\ p_1 & p_2 & \cdots & p_k \end{pmatrix}$ 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

$$l(P) = \sum_{i=1}^{n} log[f(x_i, P)] = \sum_{i=1}^{n} log[\sum_{j=1}^{k} f(x_i, \lambda_j) p_j] \tag{2}$$

It is one of the important aspects of inference in mixture distributions to maximize the log-likelihood (2). Two cases have to be *clearly* distinguished. For one, the *number of mixture components* k might be fixed in advance, 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).

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*. As a major tool the directional derivative at P in the direction Q is used which is defined as: $\Phi(P, Q) = lim_{\alpha \to 0} = [l((1-\alpha)P + \alpha Q) - l(P)]/\alpha$. Note that $\Phi(P, Q)$ 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_λ : $\Phi(P, Q_\lambda) = \sum_i f(x_i, \lambda)/f(x_i, P) - n$. 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:$d(\lambda, P) = \frac{1}{n}\sum_{i=1}^{n} \frac{f(x_i,\lambda)}{f(x_i,P)}$ 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. The gradient function also serves as a tool for creating *globally convergent* algorithms like the Vertex-Direction Method or the Vertex-Exchange-Method among others (see Böhning (2000) for details).

The Problem of Multiple Maxima When Number of Components Is Fixed

The algorithms for the flexible number of components case will deliver some estimate $\hat{k}$ of the number of components k. Frequently however, it is desired to keep the number of components k *fixed*. The reasons for doing so might be manifold, one of them is mentioned as follows. 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). 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 et al. (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. 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 1. All studies are community-randomized trials from South-Asia or Sout-East Asia, besides the second study which is from Northern Sudan.

Table 1: *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*

Following Laird (1978) we model the effect measure $x_i = log\hat{RR}_i$ for the $i-$th study as a mixture of normal densities $f(x_i) = p_1 f(x_i, \lambda_1) + ... + p_k f(x_i, \lambda_k)$, 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 1.

In Table 2 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 $BIC = 2l(P^\infty) - (2k-1)log(n)$ 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 2 (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 2: *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	# 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

Number of Components Fixed: Globally Convergent Algorithms

Typically, the EM algorithm is employed *not* using knowledge from the existing global maximization theory. The global optimization theory is reviewed in Böhning (2000) in detail which lays ground for some simple, theory-guided adjustments of the EM algorithm to circumvent local maxima. The idea is simply to combine the use of 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 for EM algorithm having exactly k components.

Step 1. Use EM algorithm to provide at convergence $P_{EM} = P^\infty$.

Step 2. Determine λ_{max} to maximize $d(\lambda, P_{EM})$ in λ.

Step 3. Determine λ_{min} such that $l(P_{EM} + P_{EM}(\lambda_{min})[Q_{\lambda_{max}} - Q_{\lambda_{min}}])$ is largest in the set with k elements

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

where $\lambda_1, \cdots, \lambda_k$ receive positive weight in P_{EM}.

Step 4. Let $P = P_{EM} + P_{EM}(\lambda_{min})[Q_{\lambda_{max}} - Q_{\lambda_{min}}]$ (Exchange λ_{max} with λ_{min}). If $l(P) > l(P_{EM})$, go to Step 1; otherwise stop.

Note that the exchnage of the two component parameters in Step 3 keep the number of components to be always exactly k. Of course, by construction, there is guarantee of *monotonicity* and, thus by, convergence. Any sequence provided by the EMGFU will converge, though there is no guarantee of convergence to a *global* maximum. However, empirical evidence and experience demonstrates grounds that this simple adjustment of the EM algorithm by means of the gradient function provides a considerable improvement. For any set of initial values given in Table 2 for $k = 2$, the EMGFU converges to the maximal log-likelihood -2.73066.

The EMGFU with Dimension Adjustment

Sometimes, however, the EM algorithm has the feature to merge two (or more) components. For example, the algorithm might have started with 3 different components, but at convergence two component iterates are identical. Thus, it might happen that the number of components is reduced from k to $k-1$ which we call the *reduction problem*. The solution of this reduction problem is provided by a *dimension adjustment step* in the EMGFU algorithm.

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 $P_{EM} = P^{\infty}$. Step 1.1 If the number of components $= k$, go to step 2. *Dimension Adjuster:* Step 1.2 If the number of components $= k-1$, determine λ_{max} to maxmimize $d(\lambda, P_{EM})$ in λ and set $P = (1-\alpha_{max})P_{EM} + \alpha_{max}Q_{\lambda_{max}}$ and go to step 1. Here α_{max} is chosen to maximize the likelihood on the line connecting P_{EM} with $Q_{\lambda_{max}}$ (see Böhning 2000). If $\alpha_{max} > 0$ with $l(P) > l(P_{EM})$ 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.

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.

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.

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.

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.