ELSEVIER

Journal of Statistical Planning and Inference 47 (1995) 5–28

journal of statistical planning and inference

A review of reliable maximum likelihood algorithms for semiparametric mixture models☆

Dankmar Böhning

Department of Epidemiology, Free University Berlin, Augustastr. 37, 12203 Berlin, Germany

Abstract

This contribution reviews some of the major developments in the algorithmic area connected with the construction of maximum likelihood estimators in semiparametric mixture models. Mixture models arise in a natural way in that they are modelling unobserved population heterogeneity. It is assumed that the population consists of a possibly unknown number k of subpopulations with parameters $\vartheta_1, \ldots, \vartheta_k$ receiving weights $\alpha_1, \ldots, \alpha_k$. Since it is not possible to observe the subpopulation membership, one can only observe data coming from the marginal or mixture density $f(x, P) = \sum_{j=1}^{k} f(x, \vartheta_j)\alpha_j$ with P giving mass α_j to ϑ_j. The log-likelihood $\lambda(P)$ becomes $\sum_i \log f(x_i, P)$ and it is easily seen that $\lambda(P)$ forms a concave functional on the convex set of all probability distributions and this property in essence produces the basis for the construction principles of reliable maximum likelihood algorithms. In Section 1, the gradient function is discussed. Any optimization algorithm needs a search direction and the gradient function serves as a basis for finding vertex directions. Section 2 discusses vertex direction methods. The fundamental idea of vertex direction methods is as follows. Consider a vertex P_ϑ (the probability measure putting all mass at ϑ) and set up the convex combination of some current P and P_ϑ: $(1-\alpha)P + \alpha P_\vartheta$, where α is called the step-length. Good choices will be discussed in the next section. The gradient function is used to find the appropriate vertex direction. If ϑ is chosen to maximize the gradient function the associated convex combination is called the *vertex direction method* (VDM). Large improvements of this method such as the *vertex exchange method* (VEM) or the *intra simplex direction method* (ISDM) will be discussed. In Section 3, the problem of finding a monotonic step-length given some current value P and a search direction H is studied. The idea of estimating the area above the second derivative curve is introduced and related to existing algorithms. The concept of area overestimation then leads to the monotonicity of the associated algorithm. Section 4 discusses the problem of maximizing a concave function of a finite number of probabilities. Three algorithmic approaches are discussed: the projection approach, the transformation approach and the generalized EM approach. Concluding in Section 5, the case of a fixed (known) number of components in the mixture model as well as software packages available for all mentioned algorithms are discussed.

☆ Invited paper presented at the 8th Workshop on Statistical Modelling, Leuven, 5–9 July, 1993.

SSDI 0378-3758(94)00119-7

0. Introduction

The mixture model arises as a simple and natural way to model population heterogeneity. Suppose the population consists of k homogeneous subgroups or component populations (simply called *components*). A simple parametric model, such as the Poisson model, is then assumed to hold in each component. Formally, let $f(x, \vartheta_j)$ be the probability density for observation X, when sampled from the jth component. Suppose further the jth component is a fraction α_j of the total population, with $\alpha_1 + \alpha_2 + \cdots + \alpha_k = 1$. If the component membership is known, one can find maximum likelihood estimates of ϑ_j and α_j for simple models in a direct way (see Table 1).

Assuming that one samples from the entire population, *without knowledge of component membership*, then the observation X has *mixture* or *marginal* density

$$f(x, P) = \sum_{j=1}^{k} f(x, \vartheta_j)\alpha_j = \int_{\Theta} f(x, \vartheta)\, P(\mathrm{d}\vartheta),$$

where the unknown parameter vector P consists of k component parameters $\vartheta_1, \ldots, \vartheta_k$ and k component proportions $\alpha_1, \ldots, \alpha_k$. In the mathematical analysis of such a model, it is often useful to associate the unknown parameters P with a discrete probability distribution (which we also denote by P and call the *mixing distribution*) giving mass α_j to ϑ_j. Viewed in this fashion, the *number of components*, k, equals the *support* size of the discrete distribution P (i.e., the number of ϑ_j with strictly positive mass p_j). In constructing a mixture model, one must choose regarding the number of components k. We can either treat k as fixed and known, and call it the *fixed support* size case, or we can treat k itself as an *unknown* parameter, and call it the *flexible support size* case. In the latter case, we can then think of P as a completely unknown discrete distribution on the values ϑ. It is this latter case that will be discussed mainly in this paper.

Example. The following data have been discussed as well in Böhning et al. (1992). It refers to a cohort study in northeast Thailand, where the health status of 602 preschool children was checked every 2 weeks from June 1982 until September 1985. For each child it was recorded whether the child had one of the symptoms fever or

Table 1
Population heterogeneity: k subpopulations

Means	ϑ_1	$\vartheta_2 \cdots \vartheta_k$
Weights	α_1	$\alpha_2 \cdots \vartheta_k$
	$\downarrow$	$\cdots \downarrow$
	$\hat{\vartheta}_1$	$\cdots \hat{\vartheta}_k$
	$\hat{p}_1$	$\cdots \hat{p}_k$

Table 2
Distribution of the counting variable *number of illness spells* for a cohort sample of 602 preschool children in northeast Thailand

No. of illness spells	0	1	2	3	4	5	6	7	8	9	10	11
Frequency	120	64	69	72	54	35	36	25	25	19	18	18
No. of illness spells	12	13	14	15	16	17	18	19	20	21	23	24
Frequency	13	4	3	6	6	5	1	3	1	2	1	2

cough, or both together. The frequencies of these illness spells during the study period were recorded, as shown in Table 2. Fig. 1(a) shows the distribution of illness spells (denoted with solid dots in the figure).

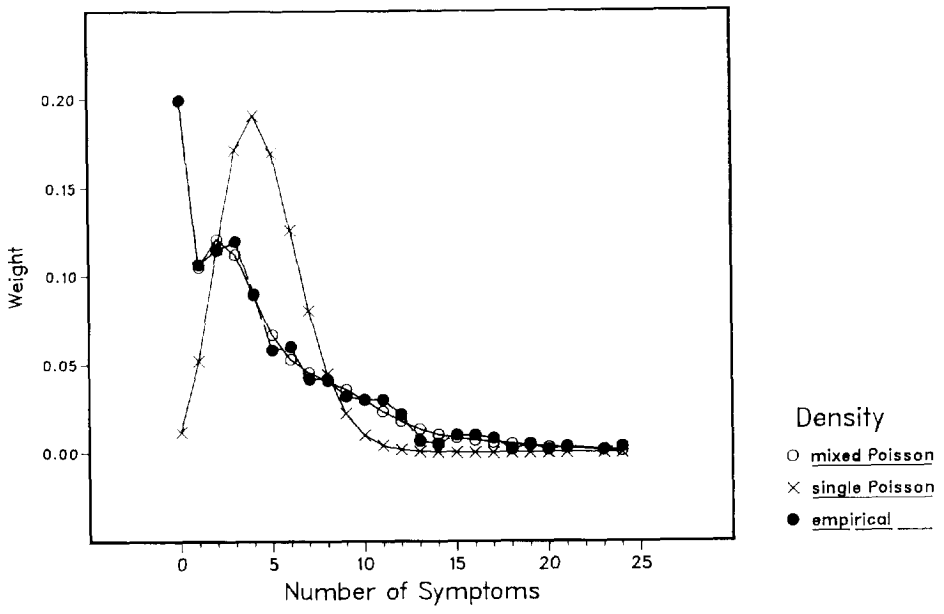
It is quite common to model this kind of count data with a Poisson distribution. The single Poisson distribution $f(x, \vartheta) = \mathrm{Po}(x, \vartheta) = \mathrm{e}^{-\vartheta}\vartheta^x/x!$ does not fit the empirical distribution very well (the crosses in Fig. 1(a) correspond to $f(x_i, \bar{x})$, sometimes also called the *fitted* or *predicted* values). Shown in the same figure are the fitted values from the Poisson mixture model

$$\mathrm{Po}(x, P) = \mathrm{Po}(x, \vartheta_1)\alpha_1 + \mathrm{Po}(x, \vartheta_2)\alpha_2 + \mathrm{Po}(x, \vartheta_3)\alpha_3,$$

which provides a much better fit. An analysis suggests (for the semiparametric maximum likelihood estimator see Fig. 1(b)) that the population consists of three components with estimates $\hat{\vartheta}_1$, $\hat{\vartheta}_2$, $\hat{\vartheta}_3$. Component 1 consists of all those children who were rarely ill; component 2, a group that tended to be sick several times (an average of 8–9 times); and component 3, a group that tended to be sick quite often (an average of 16–17 times). Note in Fig. 1(b) that the fit of the three-component mixture model coincides with the empirical fit at the origin.

The method of estimation used here is maximum likelihood. A maximum likelihood estimator $\hat{P}$ of P is defined as a probability measure $\hat{P}$ that maximizes the log-likelihood function $\lambda(P) = (1/n)\sum_{i=1}^{n} \log f(x_i, P)$. Sometimes it is more convenient to write the log-likelihood function in the form $\lambda(P) = \sum_x w_x \log f(x, P)$, with $w_x = (\text{no. of values in sample} = x)/n$. For fixed support size (i.e., k-component models), P is allowed to vary in the set Ω_k of all discrete probability measures with maximum support size k. In the case of flexible support size (i.e., k unknown), P varies in the set Ω of all probability measures. In the latter case, the estimated probability measure $\hat{P}$ is known as the *semi- or nonparametric maximum likelihood estimator* (SMLE) of the mixing distribution (Laird, 1978, 1982). For a general introduction into the topic the reader is advised to check the books of Titterington et al. (1985), McLachlan and Basford (1988), or the older one by Everitt and Hand (1981).

The main question which will be dealt with in this paper is as follows: Is there unobserved heterogeneity, and if so, how can it be estimated and the estimator algorithmically obtained?



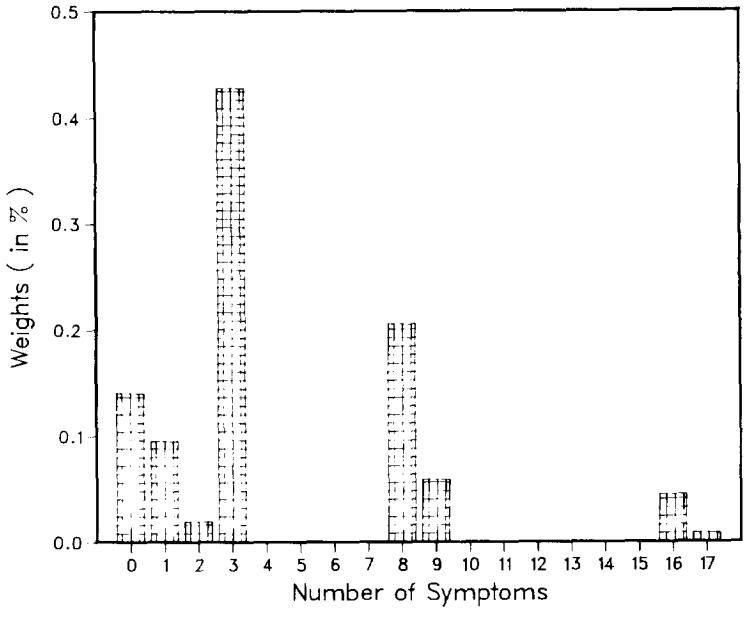


Fig. 1. (a) Empirical distribution, single and mixed Poisson of illness spells of a sample of 602 preschool children in the northeast of Thailand. (b) Semiparametric estimation of heterogeneity for a sample of 602 preschool children in the northeast of Thailand.

1. Gradient function and vertex directions

Characterizing the semiparametric maximum likelihood estimate: We consider the following:

Problem. Given a sample $x_1, \ldots, x_n$, find the SMLE for P where $\hat{P}$ is SMLE by definition if $\hat{P}$ maximizes the log-likelihood

$$\lambda(P) = \sum_x w_x \log f(x, P).$$

We do have the following properties (Lindsay, 1981, 1983a, b; Böhning, 1982; Jewell, 1982). Note that these results do not require any special assumption on the density $f(x, \vartheta)$ under consideration.

Properties. (1) λ concave in Ω, the set of all probability measures on the parameter space Θ.

(2) The directional derivative is defined as (and exists)

$$\Phi(P, Q) = \lim_{\alpha \to 0} [\lambda((1-\alpha)P + \alpha Q) - \lambda(P)]/\alpha$$
$$= \sum_x w_x \frac{f(x, Q) - f(x, P)}{f(x, P)}$$

for any two probability measures P and Q.

(3) In particular, if P_ϑ is the one point measure putting all its mass at ϑ

$$D_P(\vartheta) = \Phi(P, P_\vartheta) = \sum_x w_x \frac{f(x, \vartheta) - f(x, P)}{f(x, P)}$$

is called the *gradient function*.

(4) *General mixture maximum likelihood theorem*:

(a) $\hat{P}$ MLE $\Leftrightarrow D_{\hat{P}}(\vartheta) \leqslant 0$ for all ϑ,

(b) $D_{\hat{P}}(\vartheta) = 0$ for every support point ϑ of $\hat{P}$.

Remarks. One of the nice features of this theorem lies in the fact that a characterization of the SMLE is provided by just looking into vertex directions. Part (b) of the theorem can also be thought of as a defining equation for the weights given the support points. There are also other forms of characterizing the SMLE such as one given by Gribik and Kortanek (1971, 1977) which states that $\hat{P}$ is SMLE if and only if $\Phi(Q, \hat{P}) \geqslant 0$ for all probability measures Q. This theorem can be topographically interpreted as 'the top of the mountain is characterized by the fact that from every point on this mountain it goes upward into the direction of the top'.

The gradient function is useful in many respects. A recent approach (Lindsay and Roeder, 1992a) focuses on the graphical analysis of the residuals (observed − expected)/expected as a diagnostic tool and establishes a close relationship to the gradient function as a *smoothed* version of the residuals. In Lindsay and Roeder (1992b) question of uniqueness and identifiability in mixture models are discussed.

Computational strategy for detecting homogeneous populations: Let us suppose that the sample is coming from a homogeneous population described by some scalar mean ϑ_0. Then, in standard situations the sample mean $\bar{x}$ would be an estimate of ϑ_0 and if $D_{\bar{x}}(\vartheta) \leqslant 0$ for all ϑ, we know by the general mixture maximum likelihood theorem that $\bar{x}$ would be also the SMLE. The *critical point here* is as follows: under regular conditions we know that $l = 2n[\lambda(\hat{P}) - \lambda(\bar{x})]$ has a limiting χ^2-distribution under the null hypothesis of homogeneity with degrees of freedom equal to the difference in parameters between alternative and null hypothesis, in this case, k means $+ (k-1)$ weights $= 2k-1$ parameters under the alternative, meaning $2(k-1)$ parameters as difference. This would imply that we can expect $l = 0$ with no positive probability, even if the population is homogeneous. However, this result does not hold here, since the null hypothesis is lying in the boundary of the alternative, and so there exists a positive probability that $l = 0$ (for this point see Titterington et al. (1985) and Böhning et al. (1994)). This analysis indicates that there is a good chance to detect a homogeneous population via the general mixture maximum likelihood theorem.

Connection to optimal design theory: Many results in semiparametric mixture models have analogous counterparts in optimal design theory. The latter context can be described as follows. Given a dependent variable Y, a vector of regressors $\boldsymbol{x}^{\mathrm{T}} = (x_1, \ldots, x_p)^{\mathrm{T}}$ and a connecting linear regression model $E(y) = \beta^{\mathrm{T}}\boldsymbol{x}$, based on a sample of size N the *best linear unbiased estimator of* β is given by $\hat{\beta} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$, where $\boldsymbol{X}$ is the design matrix

$$\begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_N^{\mathrm{T}} \end{pmatrix}$$

and $\boldsymbol{Y}^{\mathrm{T}} = (y_1, \ldots, y_N)^{\mathrm{T}}$. The covariance matrix of $\hat{\beta}$ is proportional to $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ and $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ can be written as $(1/N)\sum_{i=1}^{N} \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} p_i$, $p_i = 1/N$. In planning an *optimal experiment of size N* one would like to choose those design points which minimize in some sense the covariance matrix $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$. A frequently used optimality criterion is the determinant, since it measures the contents of the dispersion ellipsoid. In other words, one wants to find that design which maximizes $\det(\sum_{i=1}^{N} \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} p_i)$, under the restriction $p_i = 1/N$ (exact design). To simplify the optimization task one gives up the restriction $p_i = 1/N$ and maximizes the determinant of $\sum_{i=1}^{k} \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} p_i$ under the only restriction $p_i \geqslant 0$ for all $i = 1, \ldots, k$ and $p_1 + \cdots + p_k = 1$. The results available in optimal design theory can be split into three parts: the *equivalence theory* with the key result in the general equivalence theorem going back to Kiefer and Wolfowitz (1960) stating the equivalence of a continuous design giving mass $p_1, \ldots, p_k$ to design points

$x_1, \ldots, x_k$ which maximizes the determinant and the design which minimizes $\max_x x^T (\sum_{i=1}^k x_i x_i^T p_i)^{-1} x$. This result compares closely to the general mixture maximum likelihood theorem (see Atwood, 1973, 1976, 1980; Böhning, 1985; Titterington, 1975, 1980; Silvey, 1980). The second part is the *duality theory* developed by Silvey and Titterington (1973) and Titterington (1975) for which the mixture model counterparts have been pointed out by Lindsay (1983) and Lesperance and Kalbfleisch (1992). The third part of development in optimal design took place in *algorithms*. Most of their mixture model counterparts will be discussed in this paper. Original contributions in the algorithmic area in optimal design are connected with the names of Atwood (1973, 1976, 1980), Böhning (1982, 1985), Gribik and Kortanek (1975, 1977), Silvey et al. (1978), Simar (1976), Titterington (1976), Torsney (1981), Tsay (1974, 1976, 1977), Wu (1978, 1983) and Fedorov (1972).

2. Vertex direction algorithms

2.1. The vertex direction method (VDM)

The vertex direction method (VDM) is based on the following property of the gradient function $D_P(\vartheta)$. If $D_P(\vartheta) > 0$ for some ϑ, then the likelihood can be increased over $\lambda(P)$ by using a distribution with some additional mass at ϑ; formally, there exists an α such that $\lambda((1-\alpha)P + \alpha P_\vartheta) > \lambda(P)$. We would like to make

$$\lambda((1-\alpha)P + \alpha P_\vartheta) - \lambda(P)$$

as large as possible. Based on a first-order approximation we have

$$\lambda((1-\alpha)P + \alpha P_\vartheta) - \lambda(P) \approx \alpha D_P(\vartheta). \tag{2.1}$$

Clearly, the right-hand side of (2.1) is maximized if we make $D_P(\vartheta)$ as large as possible. This leads to the following algorithm.

VDM

(i) Find $\vartheta_{\max}$ with $D_P(\vartheta_{\max}) = \sup_\vartheta D_P(\vartheta)$.

(ii) Set $P_{\text{new}} = (1-\alpha)P + \alpha P_{\vartheta_{\max}}$ with α monotone in the sense $\lambda(P_{\text{new}}) \geqslant \lambda(P)$.

The VDM is discussed in Wu (1978a, b), Lindsay (1983) and Böhning (1982, 1985). The question of the monotonic step-length choice will be discussed in Section 3. We have the following.

Properties. (a) *Any* sequence (P_i) created by VDM with *arbitrary* initial value meets: $\lambda(P_i) \to \lambda(\hat{P})$ monotonically.

(b) The VDM is stable, very slow and wasting energy.

Ideas on improvement: We would like to make $\lambda((1-\alpha)P + \alpha P_\vartheta) - \lambda(P)$ as large as possible. It is therefore appropriate to use a second-order approximation for this increment leading to

$$\alpha D_P(\vartheta) + \tfrac{1}{2}\alpha^2 D_P^{(2)}(\vartheta), \tag{2.2}$$

where

$$D_P^{(2)}(\vartheta) = \frac{\partial^2}{(\partial\alpha)^2}\lambda((1-\alpha)P + \alpha P_\vartheta)|_{\alpha=0}$$

is the second derivative of $\lambda((1-\alpha)P + \alpha P_\vartheta)$ with respect to α computed at $\alpha = 0$ and becomes equal to

$$-\sum_x w_x \left(\frac{f(x,\vartheta) - f(x,P)}{f(x,P)}\right)^2.$$

If we replace α in (2.2) by its maximizing value $\alpha_{\max} = -D_P(\vartheta)/D_P^{(2)}(\vartheta)$ we obtain (for $\alpha = \alpha_{\max}$)

$$(2.2) = -\tfrac{1}{2} D_P(\vartheta)^2 / D_P^{(2)}(\vartheta) \equiv \Delta_P(\vartheta).$$

A modification of the VDM would use the function $\Delta_P(\vartheta)$ in step (i) instead of $D_P(\vartheta)$.

Figs. 2 and 3 show $\max_\alpha \lambda((1-\alpha)P + P_\vartheta) - \lambda(P)$ with both approximations $D_P(\vartheta)$ (*linear*) and $\Delta_P(\vartheta)$ (*quadratic*) for two configurations. Both configurations consider the data of Example 1. In Fig. 2 the current P gives equal weight 1/5 to 0, 1, 5, 10, 15 whereas in Fig. 3 the current P gives equal weight 1/4 to 1, 5, 10, 15. In both cases, the quadratic approximation is evidently much better. However, whereas in Fig. 2 the VDM and the modified VDM would choose similar vertex directions since the values which maximize $D_P(\vartheta)$ and $\Delta_P(\vartheta)$ are rather similar, in Fig. 3 the modified VDM would choose a completely different one, as the ϑ-value for which $D_P(\vartheta)$ becomes largest is quite different to the one which maximizes $\Delta_P(\vartheta)$.

2.2. *The vertex exchange method (VEM)*

The vertex exchange method (VEM) is based on the following idea. If ϑ maximizes $D_P(\vartheta)$ and ϑ^* minimizes $D_P(\vartheta)$ in the support of P, then we would seem to need more mass at ϑ and less at ϑ^*. Formally, $P_{\text{new}} = P + \alpha p^*[P_\vartheta - P_{\vartheta^*}]$ realizes this idea of moving mass from ϑ^* to ϑ; at $\alpha = 0$ we just have $P_{\text{new}} = P$, at $\alpha = 1$ we have $P_{\text{new}} = P + p^*[P_\vartheta - P_{\vartheta^*}]$, meaning the 'bad' support point ϑ^* is replaced by ϑ. Note that $p^* = P(\vartheta^*)$. This latter exchange process has been the basis for the name *vertex exchange algorithm*. Formally, the choice of ϑ and ϑ^* can be motivated by a first-order approximation of the increment

$$\lambda(P + \alpha p^*[P_\vartheta - P_{\vartheta^*}]) - \lambda(P) \approx \alpha p^*(D_p(\vartheta) - D_p(\vartheta^*))$$

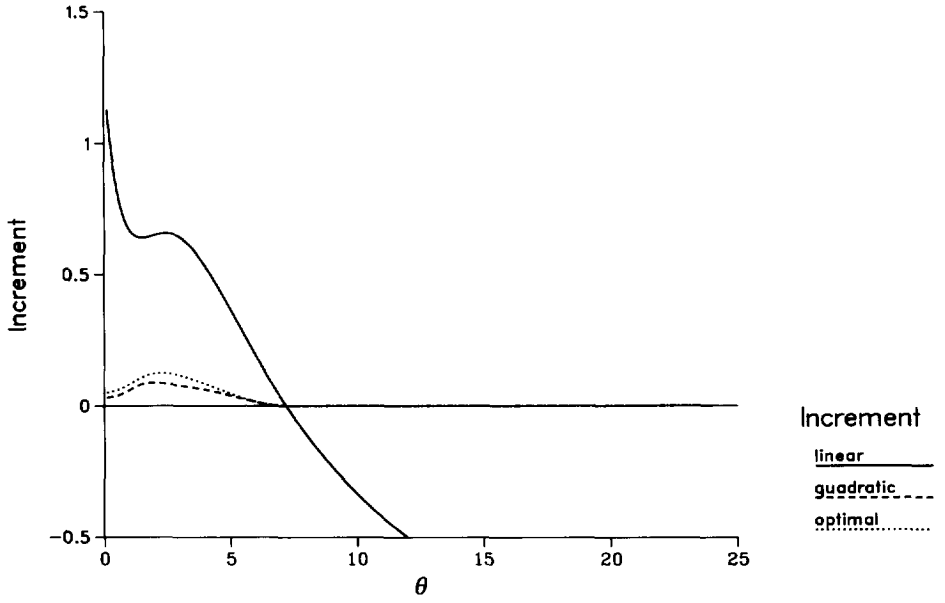


Fig. 2. Optimal increment $\lambda((1-\alpha)P+\alpha P_\theta)-\lambda(P)$ with respect to α and linear and quadratic approximation; current value P gives equal mass to 0, 1, 5, 10, 15, data are the 'illness spells' from Example 1.

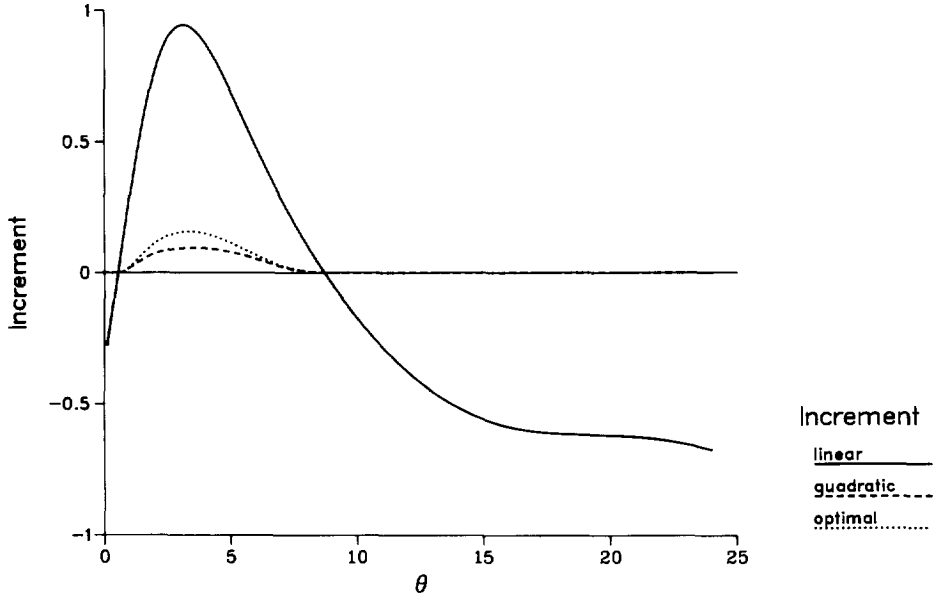


Fig. 3. Optimal increment $\lambda((1-\alpha)P+\alpha P_\theta)-\lambda(P)$ with respect to α and linear and quadratic approximation; current value P gives equal mass to 1, 5, 10, 15, data are the 'illness spells' from Example 1.

which we would like to make as large as possible. Consequently, one would like to maximize $D_P(\vartheta)$ in ϑ over the whole parameter interval and to minimize $D_P(\vartheta)$ in the support of P.

VEM

(i) Find $\vartheta_{\max}$ with $D_P(\vartheta_{\max}) = \sup_\vartheta D_P(\vartheta)$.
(i') Find $\vartheta_{\min}$ with $D_P(\vartheta_{\min}) = \sup_{\vartheta^* \in \operatorname{supp}(P)} D_P(\vartheta^*)$.
(ii) Set $P_{\text{new}} = P + \alpha p_{\min}[P_{\vartheta_{\max}} - P_{\vartheta_{\min}}]$ with α monotone in the sense $\lambda(P_{\text{new}}) \geqslant \lambda(P)$.

For a detailed discussion of the VEM the reader may look at Böhning (1985, 1986).

Properties. (a) *Any* sequence (P_i) created by the VEM with *arbitrary* initial value meets: $\lambda(P_i) \to \lambda(\hat{P})$ monotonically.

(b) The VEM is stable and converging better than the VDM. Lesperance and Kalbfleisch (1992) give an example in which the VDM needs 2177 iterations, whereas the VEM needs only 143 iterations to achieve the same stopping accuracy.

Ideas on improvements: A limited numerical experience shows that changing D_P to Δ_P in selecting $\vartheta_{\max}$ and $\vartheta_{\min}$ improves the convergence behavior, although it appears to be no 'breakthrough'. Alternatively, one can again think of a quadratic approximation in α of

$$\lambda(P + \alpha p^*[P_\vartheta - P_{\vartheta^*}]) - \lambda(P)$$

which is provided by

$$\alpha p^*(D_P(\vartheta) - D_p(\vartheta^*)) + \tfrac{1}{2}\alpha^2 p^{*2} \tilde{D}_P^{(2)}(\vartheta, \vartheta^*), \tag{2.3}$$

where

$$\tilde{D}_P^{(2)}(\vartheta, \vartheta) = \frac{\partial^2}{(\partial\alpha)^2} \lambda(P + \alpha[P_\vartheta - P_{\vartheta^*}])|_{\alpha=0}$$

$$= -\sum_x w_x \left(\frac{f(x, \vartheta) - f(x, \vartheta^*)}{f(x, P)} \right)^2.$$

If we replace α in (2.3) by its maximizing value $\alpha_{\max} = -(D_P(\vartheta) - D_P(\vartheta))/[p^* \tilde{D}_P^{(2)}(\vartheta, \vartheta^*)]$ we find that (for $\alpha = \alpha_{\max}$)

$$(2.3) = -\tfrac{1}{2}(D_P(\vartheta) - D_P(\vartheta^*))/\tilde{D}_P^{(2)}(\vartheta, \vartheta^*)$$

$$\equiv \Delta_P(\vartheta, \vartheta^*).$$

A computational strategy could be now to find ϑ^* to minimize $D_P(\vartheta^*)$ (or $\Delta(\vartheta^*)$) in $\operatorname{supp}(P)$ and then, in a second step, to find ϑ to maximize $\Delta(\vartheta, \vartheta^*)$ in ϑ. With this modification, the VEM would need 376 iterations (466 iterations with $\Delta_P(\vartheta^*)$ instead of $D_P(\vartheta^*)$ in the first step), in contrast to 1927 iterations for the traditional VEM. Computation time was 7 s (8 s) for the modified VEM, and 14 s for the traditional VEM.

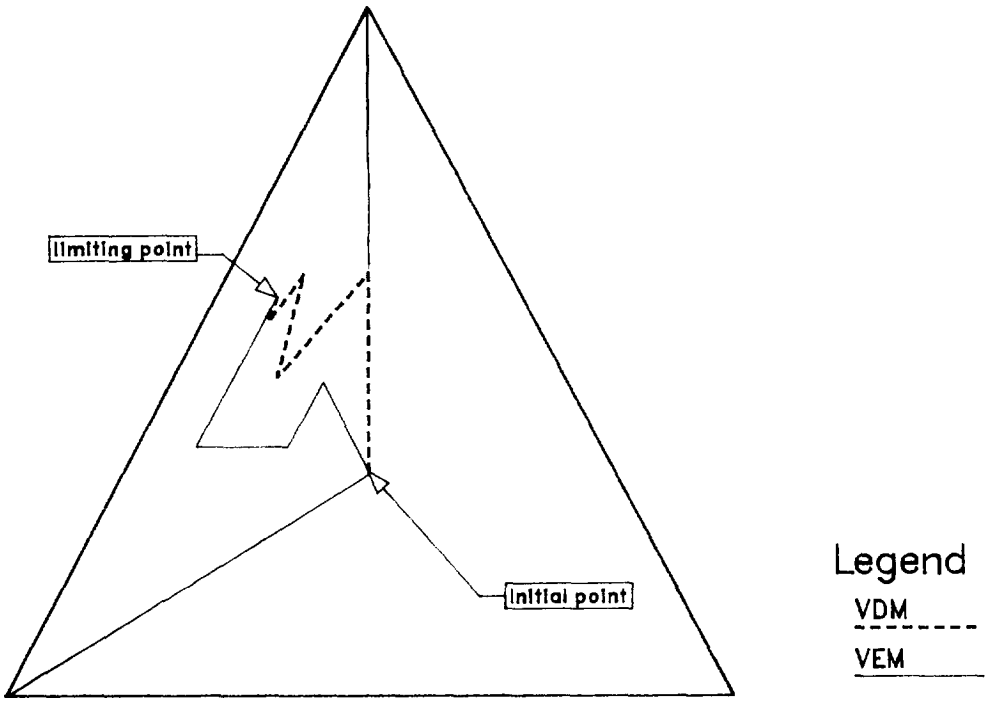


Fig. 4. VDM, VEM and ISDM for two-dimensional simplex.

2.3. *ISDM (intra simplex direction method)*

The intra simplex direction method (ISDM) has been suggested by Lesperance and Kalbfleisch (1992). The method works as follows.

ISDM

(i) Find all 'local' maxima $\vartheta_1, \ldots, \vartheta_m$ of $D_P(\vartheta)$!.

(ii) Maximize $\lambda((1-\alpha_0)P + \sum_{i=1}^{m} \alpha_i P_{\vartheta_i})$ in $\alpha_0, \ldots, \alpha_m$ *subject to* $\alpha_i \geqslant 0$ and $\alpha_0 + \cdots + \alpha_m = 1$.

(iii) Set $P_{\text{new}} = (1-\alpha_0)P + \sum_{i=1}^{m} \alpha_i P_{\vartheta_i}$.

Properties. (a) *Any* sequence (P_i) created by ISDM with *arbitrary* initial value meets: $\lambda(P_i) \to \lambda(\hat{P})$ monotonically.

(b) Lesperance and Kalbfleisch (1992) point out that the ISDM is stable and very fast. In the example mentioned before, the ISDM needed 11 iterations in contrast to 2177 for the VDM and 143 for the VEM. However, one has to keep in mind that there is an increased complexity involved in solving step (ii). We will consider algorithms to solve this substep in Section 4. Fig. 4 shows the differences in three methods. The VDM moves always towards the direction of a vertex and the VEM moves parallel to

the edges of the simplex. The ISDM would work in subsimplex shown on the left side of Fig. 4 and find the solution in one step in this case.

(c) The ISDM is storage friendly, since only the vector $\{f(x, P) | w_x > 0\}$ needs to be stored (and *not* P itself), and is, in this respect, similar to the VDM.

Ideas on improvement: The analysis provided so far suggests that it might be promising to use $A_P(\vartheta)$ in step (i) instead of $D_P(\vartheta)$.

2.4. Further concepts

The EM algorithm (for number of components known) has been discussed extensively for mixtures (Redner, 1980; Redner and Walker, 1984; Hathaway, 1983, 1986; Lindsay, 1984; Wu, 1983). Based on this algorithm we can develop the following:

Computational strategy:

(i) (EM step for support size k)

$$\begin{array}{ll} \vartheta_1 \longrightarrow \vartheta_1^{\mathrm{EM}}, & \alpha_1 \longrightarrow \alpha_1^{\mathrm{EM}} \\ \vdots & \vdots \\ \vartheta_k \longrightarrow \vartheta_k^{\mathrm{EM}}, & \alpha_k \longrightarrow \alpha_k^{\mathrm{EM}} \end{array}$$

work for a while here, then go to step (ii)!

(ii) (Increase support size)

Find ϑ to maximize $D_P(\vartheta)$.

Set $\vartheta_{k+1} = \vartheta$, $k = k + 1$, go to step (i)!

This strategy has been developed by DerSimonian (1986, 1990).

Other ideas use the fact that there is a dual problem connected with the mixture problem (primal problem) (Lindsay, 1983; Böhning, 1983). Lesperance and Kalbfleisch (1992) use the dual problem to apply algorithms developed in semi-infinite programming (Coope and Watson, 1985). One could also use the primal–dual relationship to develop an algorithm that is based on projecting back and forth from primal to dual.

3. A class of monotonic step-length estimators

In this section we provide a class of step-length estimators which meet the goal of making the step-length monotonic in the sense of increasing the likelihood at each step. There are several ways to construct a monotonic algorithm; here, we review the relevant concepts from Böhning and Lindsay (1988) and Böhning (1989). In the case of interest at hand we wish to maximize over α a log-likelihood of the form

$$\varphi(\alpha) = \lambda(P + \alpha H) - \lambda(P),$$

where typical examples of the direction H include the VDM direction $H = (P_\vartheta - P)$ or the VEM direction $H = p^*(P_\vartheta - P_{\vartheta^*})$. We observe that

$$\varphi(\alpha) = \sum_x w_x \log f(x, P + \alpha H)$$

$$= \sum_x w_x \log [f(x, P) + \alpha f(x, H)]$$

$$= \sum_x w_x \log (A_x + \alpha B_x)$$

with the first four derivatives equal to

$$\varphi'(\alpha) = \sum_x w_x \frac{B_x}{A_x + \alpha B_x}, \qquad \varphi''(\alpha) = -\sum_x w_x \left(\frac{B_x}{A_x + \alpha B_x}\right)^2 \leqslant 0,$$

$$\varphi'''(\alpha) = 2\sum_x w_x \left(\frac{B_x}{A_x + \alpha B_x}\right)^3, \qquad \varphi^{(\mathrm{iv})}(\alpha) = -6\sum_x w_x \left(\frac{B_x}{A_x + \alpha B_x}\right)^4 \leqslant 0,$$

implying that φ and φ'' are concave. We call this the *double concavity property* of the mixture likelihood. We consider

$$\mathrm{AREA}(\alpha) = \int_0^\alpha \varphi''(\tau)\,\mathrm{d}\tau = \varphi'(\alpha) - \varphi'(0),$$

the area above φ'' *from* 0 *to* α (see Fig. 5). Specially, if α equals the optimal step-length $\hat{\alpha}$ we have

$$\mathrm{AREA}(\hat{\alpha}) = \varphi'(\hat{\alpha}) - \varphi'(0) = -\varphi'(0).$$

This equation allows an interesting perspective. Although we do not know $\hat{\alpha}$, we do know $\mathrm{AREA}(\hat{\alpha})$, the area above φ'' from 0 to $\hat{\alpha}$ which is $-\varphi'(0)$.

Algorithms differ in the way that they provide an estimate area(α) of AREA(α). Many well-known algorithms can be reproduced by equating the estimated area(α) with the 'true' AREA(α). The estimating equation is

$$\mathrm{area}(\alpha) = \mathrm{AREA}(\hat{\alpha}). \tag{3.1}$$

Let us look at two examples.

Example 1. Let $\mathrm{area}_{\mathrm{NR}}(\alpha) = \alpha\varphi''(0)$, the rectangle estimator with baseline α and height $\varphi''(0)$. Solving (3.1) leads to $\alpha\varphi''(0) = \mathrm{area}_{\mathrm{NR}}(\alpha) = \mathrm{AREA}(\hat{\alpha}) = -\varphi'(0)$, from which we get the Newton–Raphson (NR) estimator $\alpha_{\mathrm{NR}} = -\varphi'(0)/\varphi''(0)$.

Example 2. Let $\mathrm{area}_{\mathrm{sec}}(\alpha) = \alpha\,\mathrm{AREA}(1)$ $(= \varphi'(1) - \varphi'(0))$, the estimator that assumes AREA(α) can be modelled as a straight line. Solving (3.1) leads to $\alpha[\varphi'(1) - \varphi'(0)] = \mathrm{area}_{\mathrm{sec}}(\alpha) = \mathrm{AREA}(\hat{\alpha}) = -\varphi'(0)$, from which the secant estimator $\alpha_{\mathrm{sec}} = -\varphi'(0)/[\varphi'(1) - \varphi'(0)]$ is easily derived.

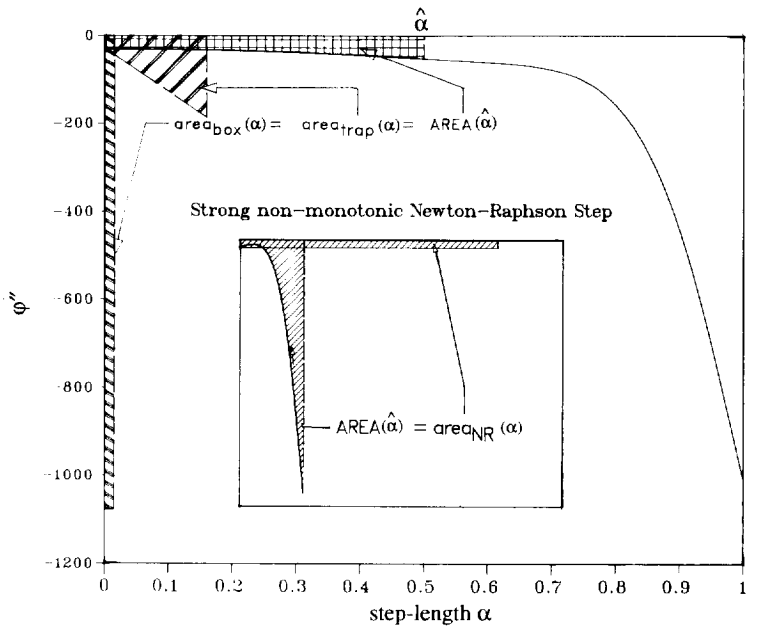


Fig. 5. Illustration of estimating Eq. (3.1) for *rectangular* and *trapezoidic* estimator of AREA($\hat{\alpha}$).

Do we obtain any new insights from the duality

area estimators	⟷	step-length algorithms	?

In fact, it is possible to clarify the monotonicity of existing algorithms, and, secondly, it can be helpful in constructing better monotonic step-length estimators. The subplot in Fig. 5 shows a strong non-monotonic, overshooting Newton–Raphson step. The geometrically obvious reason for this is that height ($\varphi''(0)$) of the rectangle is so small that a wide baseline (α_{NR}) is needed to meet the estimating equation $\text{area}_{NR}(\alpha_{NR}) = \text{AREA}(\hat{\alpha})$. To avoid this effect we need an estimator which over-estimates the true area. In Böhning (1989) the following result is proved.

Result 3.1. If $\text{area}(\alpha) \leqslant \text{AREA}(\alpha)$ for all α then a solution α^* of (3.1) is monotonic in the sense $\varphi(\alpha^*) \geqslant \varphi(0)$.

It is appropriate to construct a monotonizing version of the Newton–Raphson step by replacing the second derivative at 0 by a global lower bound.

Example 1 (*Monotonizing Newton–Raphson step*). We set $M = \inf_\alpha \varphi''(\alpha)$ and define $\text{area}_{box}(\alpha) = \alpha M \leqslant \text{AREA}(\alpha)$. Solving (3.1) leads to $\alpha_{box} = -\varphi'(0)/M$. Now we can

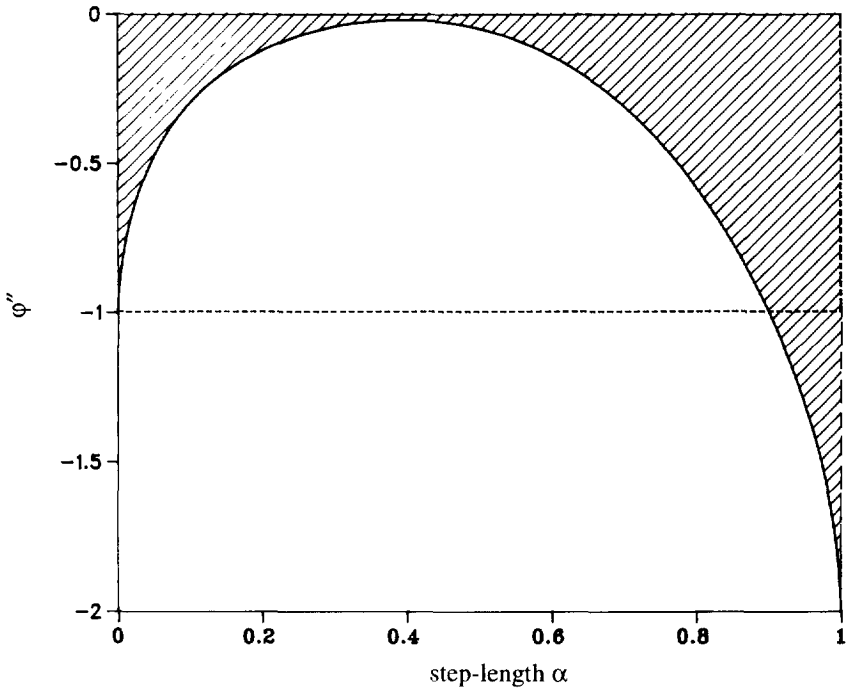


Fig. 6. Illustration of the monotonicity condition $\varphi''(0) \leqslant \text{AREA}(1)$.

exploit the double concavity property to see that the minimum M is obtained at the end-points of the interval $[0, 1]$: $M = \min\{\varphi''(0), \varphi''(1)\}$. The corresponding geometric illustrations are given in Fig. 5.

Example 2 (*Improving the box-estimator*). Clearly, the box-estimator leads to a conservative step-length choice. Why not approximate AREA(α) with trapezoid? The associated formula is

$$\text{area}_{\text{trap}}(\alpha) = \alpha\varphi''(0) + \alpha^2\{\varphi''(1) - \varphi''(0)\}/2$$

and (3.1) takes the form

$$\alpha\varphi''(0) + \alpha^2\{\varphi''(1) - \varphi''(0)\}/2 + \varphi'(0) = 0. \tag{3.2}$$

See Fig. 5. Eq. (3.2) has unique solution α_{trap} in $[0, 1]$. Note that we have the property: $0 \leqslant \alpha_{\text{box}} \leqslant \alpha_{\text{trap}} \leqslant \hat{\alpha}$.

There exists a further refinement of the Newton–Raphson step. If $\varphi''(0) \leqslant \varphi''(1)$ then area_{NR} will necessarily meet the overestimating condition of result (3.1). But this will also be true, if $\varphi''(0) \leqslant \text{AREA}(1)$, in other words if the box with base line from 0 to 1 and height $\varphi''(0)$ overestimates the area above the curve φ'' from 0 to 1. Such a situation is demonstrated in Fig. 6. Only if the latter condition is violated, we would

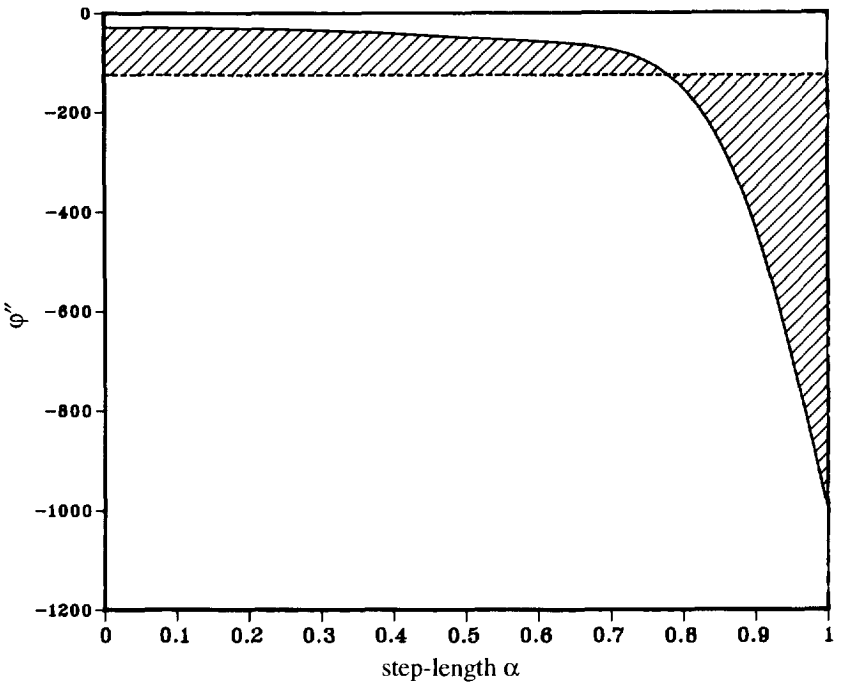


Fig. 7. Optimal choice of curvature average, shaded areas are equal.

need to modify the Newton–Raphson step. It is then appropriate to construct a box with height as an 'optimal' average of $\varphi''(0)$ and $\varphi''(1)$ in such a way that the condition $(1-\beta)\varphi''(0)+\beta\varphi''(1)=\text{AREA}(1)\ [=\varphi'(1)-\varphi'(0)]$ is met. See Fig. 7. The corresponding area estimator is $\text{area}_{\text{optimal}}(\alpha)=\alpha[(1-\beta)\varphi''(0)+\beta\varphi''(1)]=\alpha[\varphi'(1)-\varphi'(0)]=\text{area}_{\text{sec}}(\alpha)$. We summarize:

Result 3.2 (Böhning, 1989).

If	$\varphi''(0)\leqslant\varphi''(1)$ or $\varphi''(0)\leqslant\text{AREA}(1)$	α_{NR} is monotonic,
otherwise		α_{sec} is monotonic.

This result leads to a modified version of the

VEM

(i) Find $\vartheta_{\min}$ with $D_P(\vartheta_{\min})=\sup_{\vartheta^*\in\text{supp}(P)}D_P(\vartheta^*)$.

(ii) Find $\vartheta_{\max}$ with $\Delta_P(\vartheta_{\max},\vartheta_{\min})=\sup_{\vartheta}\Delta_P(\vartheta,\vartheta_{\min})$ where $\Delta_P(\vartheta,\vartheta^*)$ is defined as
  (a) $p^*(D_P(\vartheta)-D_P(\vartheta^*))+\frac{1}{2}p^{*2}\tilde{D}_P^{(2)}(\vartheta,\vartheta^*)$, if $\varphi'(1)\geqslant 0$ $(\alpha=1)$,
  (b) $\alpha p^*(D_P(\vartheta)-D_P(\vartheta^*))+\frac{1}{2}\alpha^2 p^{*2}\tilde{D}_P^{(2)}(\vartheta,\vartheta^*)$ with α chosen according to Result 3.2, if $\varphi'(1)<0$.

(iii) Set $P_{\text{new}}=P+\alpha p_{\min}[P_{\vartheta_{\max}}-P_{\vartheta_{\min}}]$, α as in step (ii).

Note that here the monotonic step-length choice has been built into the vertex choosing function $\Delta_P(\vartheta, \vartheta^*)$.

4. Algorithms for finding the weights

In Section 2 *three* algorithms (VDM, VEM, ISDM), each capable of finding the SMLE in a reliable way, are represented. However, there are at least two arguments to look at the maximum likelihood problem from the point of view of maximizing the likelihood on the *finite* dimensional simplex as described in (4.1).

Maximize the concave function

$$\lambda(\alpha_1, \ldots, \alpha_k) = \sum_x w_x \log\left(\sum_{j=1}^{k} f(x, \vartheta_j)\alpha_j \right) \tag{4.1}$$

s.t. $\alpha_i \geqslant 0$ and $\alpha_1 + \cdots + \alpha_k = 1$ (or, equivalently, $\alpha \in$ finite simplex Ω).

One reason is that problem (4.1) has already occurred as a subproblem in the ISDM algorithm in step (ii). Secondly, one can think of approximating the set of all probability measures on Θ by the set of all probability measures on an approximating set $\Theta_{\text{grid}} = \{\vartheta_1, \ldots, \vartheta_k\}$ of Θ. This would also lead to problem (4.1).

4.1. A class of search directions

Wu (1978a, b) considers a class of search directions based on the mapping

$$I(z) = I^{-1}(z) = (\mathbf{1}^T A\mathbf{1})Az - (\mathbf{1}^T Az)A\mathbf{1} \quad \text{for arbitrary } k\text{-vector } z.$$

Here $\mathbf{1} = (1, \ldots, 1)^T$, A any symmetric, positive definite matrix. $\mathbf{I}$ has the following properties:

(a) $u = I(z)$ satisfies the constraint $\mathbf{1}^T u = \sum_i u_i = 0$, which means $(\alpha + u) = 1$. $I(\nabla\lambda(\alpha))$ offers good search directions since the directional derivative

(b) $$\begin{aligned} \Phi(\alpha, I(\nabla\lambda(\alpha))) &= \nabla\lambda(\alpha)^T I(\nabla\lambda(\alpha)) \\ &= (\mathbf{1}^T A\mathbf{1})\nabla\lambda(\alpha)^T A\nabla\lambda(\alpha) - (\mathbf{1}^T A\nabla\lambda(\alpha))\nabla\lambda(\alpha)^T A\mathbf{1} \\ &\geqslant 0 \end{aligned}$$

by the Cauchy–Schwarz inequality, with equality if and only if $\alpha = \hat{\alpha}$. $I(\nabla\lambda(\alpha)) = I(\alpha)$ is direction of *ascent*.

(c) For $\alpha \in \Omega$, $\alpha + I(\alpha)$ need not be in Ω (although it is lying in the 'right' subspace, but it might be too long, see Fig. 8 for an illustration). Obviously,

$$\begin{aligned} 0 &= \alpha_i + \frac{\alpha_i}{-I_i(\alpha)} I_i(\alpha) \\ &\leqslant \alpha_i + m(\alpha) I_i(\alpha), \end{aligned}$$

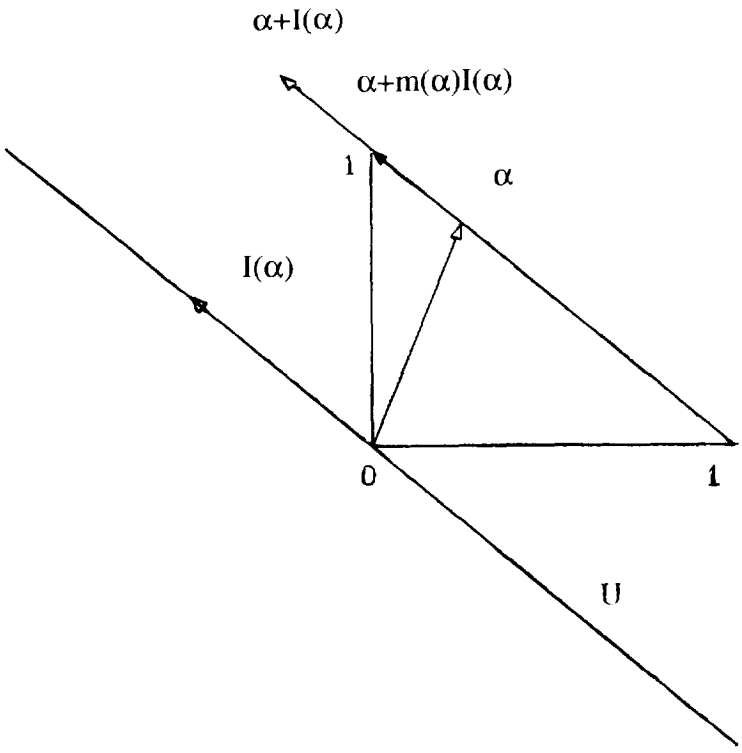


Fig. 8. Length adjustment of $I(\alpha)$.

where $m(\alpha) = \min\{-\alpha_i/I_i(\alpha) | I_i(\alpha) < 0\}$ is the appropriate length adjustment. Then $\alpha + m(\alpha) I(\alpha)$ lies in Ω. This leads to the following algorithm.

Algorithm

(i) Compute $A = A(\alpha)$, $\nabla\lambda(\alpha)$, $I(\nabla\lambda(\alpha))$.

(ii) $\alpha_{\text{new}} = \alpha + \beta m(\alpha) I(\alpha)$ with β monotonic.

Properties. The algorithm not only offers a reliable procedure for constructing the maximum likelihood estimate, but it also provides a frame to include well-known algorithms such as the projected gradient ($A = E$) or projected Newton ($A = (-\partial^2\lambda/\partial\alpha_i\partial\alpha_j)$) ones. The algorithm was suggested by Wu (1978a, b) and with A as the negative, inverted Hessian matrix by Atwood (1976, 1980), both in the context of optimal experimental design (Silvey, 1980). In Böhning and Hoffmann (1982) a detailed review of this class of algorithms is given, whereas in Böhning and Hoffmann (1986) a Fletcher–Powell type choice of the A-matrix is discussed and compared with the projected Newton method.

4.2. A transformation approach

In Böhning (1984), analogous to logistic regression, a transformation is suggested that allows the application of constraint free optimization techniques. The corresponding transformation is the

EXPIT-transformation:

$$(z_1, \ldots, z_{k-1}) \to (e^{z_1}, \ldots, e^{z_{k-1}}, 1)/\left(1 + \sum_j e^{z_j}\right) \in \Omega^+.$$

z can now vary freely in $\mathbb{R}_{k-1}$.

Define $L(z) = \lambda(\text{EXPIT}(z))$

$$= \sum_i w_i \log\left(\sum_{j=1}^{k-1} f_{ij} e^{z_j} + f_{ik}\right) - \log\left(1 + \sum_{l=1}^{k-1} e^{z_l}\right), \tag{4.2}$$

where the w_i's correspond to those $w_x > 0$ and $f_{ij} = f(x_i, \vartheta_j)$. The gradient and second derivative matrix are easy to obtain.

Gradient. The partial derivatives can be simply found as

$$\frac{\partial L}{\partial z_l} = \sum_{i=1}^{n} w_i \frac{f_{il} e^{z_l}}{\sum_{j=1}^{k-1} f_{ij} e^{z_j} + f_{ik}} - \frac{e^{z_l}}{1 + \sum_j e^{z_j}} = \sum_i w_i f_{il} \alpha_l / f(x_i, P) - \alpha_l. \tag{4.3}$$

Hessian. The element (l', l) of the second derivative matrix is given as

$$\frac{\partial^2 L}{\partial z_{l'} \partial z_l} = \sum_{i=1}^{n} w_i \frac{\delta_{ll'} f_{il} e^{z_l} (\sum_{j=1}^{k-1} f_{ij} e^{z_j} + f_{ik}) - f_{il} e^{z_l} f_{il'} e^{z_{l'}}}{(\sum_{j=1}^{k-1} f_{ij} e^{z_j} + f_{ij})^2}$$

$$- \frac{\delta_{ll'} e^{z_l} (1 + \sum_j e^{z_j}) - e^{z_l} e^{z_{l'}}}{(1 + \sum_{j=1}^{k-1} e^{z_j})^2}$$

$$= \sum_{i=1}^{n} w_i (\delta_{ll'} f_{il} \alpha_l / f(x_i, P) - f_{il} \alpha_l f_{il'} \alpha_{l'} / f(x_i, P)^2) - (\delta_{ll'} \alpha_l - \alpha_l \alpha_{l'}). \tag{4.4}$$

Here $\delta_{ll'}$ is the Kronecker symbol ($\delta_{ll'} = 1$ if $l = l'$ and $\delta_{ll'} = 0$ otherwise). Note that the gradient as well as the Hessian of (z) depend on the $(k-1)$-vector z only through α. This observation leads to the following *rescaling modification of the Newton–Raphson* algorithm.

Rescaling algorithm

Step 0 (*Initialization*). Choose weights $\alpha_1, \ldots, \alpha_k$ ($\alpha_j \geqslant 0$, $\sum_{j=1}^{k} \alpha_j = 1$) (Comment: A default value of $\alpha_j = 1/k$ corresponds to the z-vector $\mathbf{0}$).

Step (i). Compute the gradient $\nabla L(z)$ and the Hessian $\nabla^2 L(z)$ via formulae (4.3) and (4.4), respectively.

Step (ii) (*NR step*). Compute $z_{NR} = z - \nabla^2 L(z)^{-1} \nabla L(z)$.

Step (iii) (*Rescaling*). Set $z = z_{NR}$, and compute $\alpha = \text{EXPIT}(z)$, and go to step (i).

Properties. The transformation approach is studied in Böhning (1984) as well as in Formann (1980, 1982). One of its disadvantages is that it works only for $\alpha \in \Omega^+$ (all weights positive) and is therefore only useful to determine the weights *after the support points of the SMLE have been identified.*

4.3. A fixed point concept

The following concept goes back to a suggestion of Silvey et al. (1978) and Torsney (1981a, b) and is further studied in Böhning (1983). It is based on the fact that in the finite simplex case the general mixture maximum likelihood theorem states that the following four statements are equivalent:

(i) $\hat{\alpha}$ MLE.

(ii) $\dfrac{\partial \lambda}{\partial \alpha_i}(\hat{\alpha}) = \nabla \lambda(\hat{\alpha})^{\mathrm{T}} \hat{\alpha}.$

(iii) $\hat{\alpha}$ is a fixed point of F, where the ith component of

$$F_i(\alpha) = \frac{\partial \lambda}{\partial \alpha_i}(\alpha)\alpha_i / \nabla \lambda(\alpha)^{\mathrm{T}} \alpha$$

(corresponds to the EM iteration).

(iv) $\hat{\alpha}$ a fixed point of $F^{(\delta)}$, where the ith component is given as

$$F_i^{(\delta)}(\alpha) = \left[\frac{\partial \lambda}{\partial \alpha_i}(\alpha) \right]^{(\delta)} \alpha_i / \nabla^{(\delta)} \lambda(\alpha)^{\mathrm{T}} \alpha$$

with $\delta > 0$ arbitrary.

Since $F^{(1)}(\alpha)$ is the EM iteration, $F^{(\delta)}(\alpha)$ can be viewed as a *generalized* EM iteration. Fellman (1987) looks at values of δ other than 1 to speed up the iteration and finds that values around $\delta = 0.9$ give better convergence results.

5. Miscellaneous topics

5.1. Problems in the fixed components case

The preceding algorithms have the feature that the number of support points is flexible, potentially changing at every step. A very popular algorithm in which the number of support points is held fixed is the EM algorithm (Dempster et al., 1977). See also Laird (1978) for the mixture context. DerSimonian (1986, 1990) discusses the EM iteration together the VDM, and gives a FORTRAN subroutine. One advantage of the EM algorithm is its numerical simplicity together with a guaranteed monotonicity. For a detailed discussion of the EM algorithm for mixtures see Redner and Walker (1984), Hasselblad (1966, 1969), Titterington et al. (1985), Fahrmeir and Hamerle (1984) and Agha and Ibrahim (1984).

Initial values and stopping rule: It is one of the problems connected with the mixture model of restricted support size (fixed number of components) that the EM algorithm (as well as other ones) constructs only *local* solutions. One feature of the flexible support size approach is that it suggests initial values for the fixed support size case, which one hopes would lead to a global maximum at convergence. Another question is when to stop the EM algorithm. The criterion usually used for the EM algorithm is the size of the change in the likelihood or parameter estimates from the iteration to another. This is more a measure of lack of progress rather than of actual convergence.

Acceleration of the EM algorithm: Various attempts have been made to accelerate the EM iteration, which include those by Gediga and Holling (1988), Jamshidian and Jennrich (1993), Louis (1982) and Meilijson (1989). Detailed numerical evaluations will be necessary to see how fruitful these methods are in the mixture context. In some cases the parameter estimates are not of interest, but rather the value of the likelihood estimate. The following simple version of the Aitken acceleration (for details see Böhning et al., 1994) can be used.

Let the λ_i be the log-likelihood at iteration i. Then we achieve the usual Aitken acceleration in that we assume that $\lambda_{i+1} - \lambda_i \approx c(\lambda_i - \lambda_{i-1})$ which implies $\lambda_{i+1} - \lambda_i \approx c^i(\lambda_1 - \lambda_0)$. From here we get

$$\lambda^\infty \approx \lambda_0 + \sum_{i=0}^{\infty} c^i(\lambda_1 - \lambda_0) = \lambda_0 + \frac{1}{1-c}(\lambda_1 - \lambda_0)$$

which can be estimated by

$$\hat{\lambda}^\infty \approx \lambda_0 + \frac{1}{1-\hat{c}}(\lambda_1 - \lambda_0), \quad \hat{c} = \frac{\lambda_2 - \lambda_1}{\lambda_1 - \lambda_0}.$$

Note the monotonicity property: $\hat{\lambda}^\infty \geqslant \lambda_1$.

5.2. *Software*

One of the recent developments concerning software for mixture models is C.A.MAN developed by Böhning et al. (1992). It is menu-oriented and includes algorithms for the flexible as well as for the fixed support size case. The data can come from densities including normal, exponential, Poisson, binomial, and others. However, the inclusion of covariances — such as fitting models like $\sum_j \mathrm{Po}(x, \alpha_j + \beta^T z)p_j$ where mixing would here go over the intercept — is not yet available; for connected work in this area see Aitkin and Tunnicliffe Wilson (1980) or Dietz (1992). A specific software package based on the Poisson submodule of C.A.MAN is DISease MAPping developed by Schlattmann (1993) and Schlattmann and Böhning (1993). It allows the spatial analysis of rates (prevalence or incidence rates) or ratios (such as SMR) denoted by x/E through the specific Poisson mixture model:

$$\mathrm{Po}(x, E, P) = \mathrm{Po}(x, E\vartheta_1)\alpha_1 + \cdots + \mathrm{Po}(x, E\vartheta_k)\alpha_k.$$

Here the number of components k in the mixture model would correspond to the heterogeneity involved in the spatial structure. DISMAP is strongly graphics oriented and is able to put the found heterogeneous structure in an associated disease map.

Other software developments include the package MIX of MacDonald (1986) and the subroutines level work done by McLachlan and Basford (1988), DerSimonian (1986, 1990) and Agha and Ibrahim (1984).

References

Agha, M. and M.T. Ibrahim (1984). Maximum likelihood estimation of mixtures of distributions. *J. Roy. Statist. Soc. Ser. C* **33**, 327–332.

Aitkin, M. and G. Tunnaclife Wilson (1984). Mixture models, outliers, and the EM algorithm. *Technometrics* **22**, 325–332.

Atwood, C.L. (1973). Sequences converging to D-optimal designs of experiments. *Ann. Statist.* **1**, 342–352.

Atwood, C.L. (1976a). Computational considerations for convergence to an optimal design. *Proc. 1976 Conf. on Information Sciences and Systems*, Dept. of Electrical Engineering, Johns Hopkins Univ.

Atwood, C.L. (1976b). Convergent design sequences for sufficiently regular optimality criteria. *Ann. Statist.* **4**, 1124–1138.

Atwood, C.L. (1980). Convergent design sequences for sufficiently regular optimality criteria, II: singular case. *Ann. Statist.* **8**, 894–913.

Böhning, D. (1982). Convergence of Simar's algorithm for finding the MLE of a compound Poisson process. *Ann. Statist.* **10**, 1006–1008.

Böhning, D. (1983). A duality theorem with applications to statistics. *Math. Operationsforsch. Statist. Ser. Statist.* **14**, 551–557.

Böhning, D. (1984). Use of reparameterization in nonlinear optimization with applications to statistics and optimal design. *Comput. Statist. Quart.* **1**, 29–43.

Böhning, D. (1985). Numerical estimation of a probability measure. *J. Statist. Plann. Inference* **11**, 57–69.

Böhning, D. (1986). A vertex-exchange method in D-optimal design theory. *Metrika* **33**, 337–347.

Böhning, D. (1989). Likelihood inference for mixtures: geometrical and other constructions of monotone step-length algorithms. *Biometrika* **76**, 375–383.

Böhning, D., E. Dietz, R. Schaub, P. Schlattmann and B.G. Lindsay (1994). The distribution of the likelihood ratio of mixtures of densities from the one-parameter exponential family. *Ann. Inst. Statist. Math.*, **46**, 373–388.

Böhning, D. and K.-H. Hoffmann (1982). Numerical procedures for estimating probabilities. *J. Statist. Sim. Comp.* **14**, 283–293.

Böhning, D. and K.-H. Hoffmann (1986). A remark on to the numerical estimation of probabilities. *Statistics* **17**, 231–236.

Böhning, D. and B.G. Lindsay (1988). Monotonicity of quadratic approximation algorithms. *Ann. Inst. Statist. Math.* **40**, 641–663.

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

Coope, I.D. and G.A. Watson (1985). A projected Lagrangian algorithm for semi-infinite programming. *Math. Programming* **32**, 337–356.

Dempster, A.P., N.M. Laird and D.B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc.* **39**, 1–38.

DerSimonian, R. (1986). Algorithm AS 221: Maximum likeihood estimation of a mixing distribution. *J. Roy. Statist. Soc. Ser. C* **35**, 302–309.

DerSimonian, R. (1990). Correction to Algorithm AS 221 maximum likelihood estimation of a mixing distribution. *J. Roy. Statist. Soc. Ser. C* **39**, 176.

Dietz, E. (1992). Estimation of heterogeneity – a GLM approach. In: *Lecture Notes in Statistics*, Vol. 78, 66–71.

Everitt, B.S. and D.J. Hand (1981). *Finite Mixture Distributions*. Chapman and Hall. London.

Fahrmeir, L. and A. Hamerle (1984). *Multivariate Statistische Verfahren*. Walter de Gruyter, New York.

Fedorov, V.V. (1972). *Theory of Optimal Experiments*. Academic Press, New York.

Fellman, J. (1987). Some aspects of the iterative search for optimal designs. Preprint, Swedish School of Economics and Business Administration, Helsinki, Finland.

Formann, A.K. (1980). Neuere Verfahren der Parameterschätzung in der Latent-Class-Analyse. *Z. Differentielle Diagnostische Psychologie* 1 and 2, 107–116.

Formann, A.K. (1982). Linear logistic latent class analysis. *Biometrical J.* **24**, 171–190.

Gediga, G. and H. Holling (1988). On the convergence of the EM algorithm including different methods of Aitkin acceleration for finite mixture models. In: *COMPSTAT 88*. Short Communications and Posters. Physica, Wien.

Gribik, P.R. and K.O. Kortanek (1975). Equivalence theorems and cutting plane algorithms for a class of experimental design problems. Report No. 22, Carnegie Mellon Univ.

Gribik, P.R. and K.O. Kortanek (1977). Equivalence theorems and cutting plane algorithms for a class of experimental design problems. *SIAM J. Appl. Math.* **32**, 232–259.

Hasselblad, V. (1966). Estimation of parameters for a mixture of normal distributions. *Technometrics* **8**, 431–444.

Hasselblad, V. (1969). Estimation of finite mixtures of distributions from the exponential family. *J. Amer. Statist. Assoc.* **64**, 1459–1471.

Hathaway, R.J. (1983). Constrained maximum-likelihood estimation for a mixture of m univariate normal distributions. Statistics Tech. Report, 92, 62F10-2, Univ. of South Carolina, Columbia, SC.

Hathaway, R.J. (1986). Another interpretation of the EM algorithm for mixture distributions. *Statist. Probab. Lett.* **4**, 53–56.

Jamshidian, M. and R.I. Jnnrich (1993). Conjugate gradient acceleration of the EM algorithm. *J. Amer. Statist. Assoc.* **88**, 221–228.

Jewell, N.P. (1982). Mixtures of exponential distributions. *Ann. Statist.* **10**, 479–484.

Kiefer, J. and J. Wolfowitz (1960). The equivalence of two extremum problems. *Canad. J. Math.* **12**, 363–366.

Laird, N.M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc.* **73**, 805–811.

Laird, N.M. (1982). Empirical Bayes estimators using the nonparametric maximum likelihood estimate for the prior. *J. Statist. Comput. Simulation* **15**, 211–220.

Lesperance, M. and J.D. Kalbfleisch (1992). An algorithm for computing the nonparametric MLE of a mixing distribution. *J. Amer. Statist. Assoc.* **87**, 120–126.

Lindsay, B.G. (1981). Properties of the maximum likelihood estimator of a mixing distribution. In: C. Taillie et al. Eds., *Statistical Distributions in Scientific Work*. D. Reidel Publ., Dordrecht, 95–109.

Lindsay, B.G. (1983a). The geometry of mixture likelihoods, part I: a general theory. *Ann. Statist.* **11**, 783–792.

Lindsay, B.G. (1983b). The geometry of mixture likelihoods, part II: the exponential family. *Ann Statist.* **11**, 783–792.

Lindsay, B.G. (1984). Optimal EM algorithms in polynomial problems. Technical reports and reprints of Dept. of Statistics, The Pennsylvania State Univ.

Lindsay, B.G. and K. Roeder (1992a). Residual diagnostics for mixture models. *J. Amer. Statist. Assoc.* **87**, 785–794.

Lindsay, B.G. and K. Roeder (1992b). Uniqueness of estimation and identifiability in mixture models. *Canad. J. Statist.* **21**, 139–147.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44**, 226–233.

MacDonald, P.D.M. (1986). MIX: an interactive program for fitting mixtures of distributions. *Amer. Statist.* **40**, 53.

McLachlan, G.J. and K.E. Basford (1988). *Mixture Models and Applications to Clustering*. Marcel Dekker, New York.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *J. Roy. Statist. Soc. Ser. B* **51**, 127–138.

Redner, R.A. (1980). An iterative procedure for obtaining maximum likelihood estimates in a mixture model. Report SR-T1-04081, NASA Contract NAS9-14689, Texas A&M Univ., College Station, TX.

Redner, R.A. and H.F. Walker (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **24**, 195–239.

Schlattmann, P. (1993). Statistische Methoden zur Darstellung der räumlichen Verteilung von Krankheiten unter besonderer Beücksichtigung von Mischverteilungen. Inaugural-Dissertation am Fachbereich Medizinische Grundlagenfächer der Freien Universität Berlin.

Schattmann, P. and D. Böhning (1993). Mixture models and disease mapping. *Statist. Med.* **12**, 1943–1950.

Silvey, S.O. (1980). *Optimal design.* Chapman and Hall, London.

Silvey, S.D. and D.M. Titterington (1973). A geometric approach to optimal design theory. *Biometrika* **60**, 21–32.

Silvey, S.D., D.M. Titterington and B. Torsney (1978). An algorithm for optimal designs on finite design space. *Comm. Statist.* **A7**, 1379–1389.

Simar, L. (1976). Maximum likelihood estimation of a compound Poisson process. *Ann. Statist.* **4**, 1200–1209.

Titterington, D.M. (1975). Optimal design: some geometrical aspects of D-optimality. *Biometrika* **62**, 313–320.

Titterington, D.M. (1976). Algorithms for computing D-optimal designs on a finite design space. *Proc. 1976 Conf. on Information Sciences and Systems*, Dept. of Electrical Engineering, Johns Hopkins Univ., 213–216.

Titterington, D.M. (1980). Geometric approaches to design of experiment. *Math. Operationsforsch. Statist. Ser. Statist.* **11**, 151–163.

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

Torsney, B. (1981a). A moment inequality and monotonicity of an algorithm. Contribution to the 2nd Internat. Symp. on Semi-Infinite Programming and Applications, Univ. of Texas at Austin, 8–10 September.

Torsney, B. (1981b). Algorithms for a constrained optimization problem with applications in statistics and optimum design. Ph.D. Thesis, Univ. of Glasgow.

Tsay, J.Y. (1974). The iterative methods for calculating optimal experimental designs. Ph.D. Thesis, Purdue Univ.

Tasy, J.Y. (1976). On the sequential construction of D-optimal designs. *J. Amer. Statist. Assoc.* **71**, 671–674.

Tsay, J.Y. (1977). A convergence theorem in L-optimal design theory. *Ann. Statist.* **4**, 790–794.

Wu, C.F. (1978a). Some algorithmic aspects of the theory of optimal designs. *Ann. Statist.* **6**, 1286–1301.

Wu, C.F. (1978b). Some iterative procedurs for generating non-singular optimal designs. *Comm. Statist.* **A7**, 1399–1412.

Wu, C.F. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11**, 95–103.