Chapter 3

Analysis of Mixtures

3.1 Introduction

Finite mixtures of distributions have provided a mathematical based approach to the statistical modelling of a wide variety of random phenomena. Because of their usefulness as an extremely flexible method of modelling, finite mixture models have continued to receive increasing attention over the years, from both a practical and theoretical point of view. Indeed, in the past decade the applications of finite mixture models have widened considerably. Fields in which mixture models have been successfully applied include astronomy, biology, genetics, medicine, economics, engineering, and marketing among many other fields in the biological, physical and social sciences. In these applications, finite mixture models underpin a variety of techniques in major areas of statistics, including cluster and latent analysis, image analysis, and survival analysis, in addition to their more direct role in data analysis and inference of providing descriptive models for distributions.

The usefulness of mixture distributions in the modeling of heterogeneity in a cluster analysis context is obvious. Any continuous distribution can be approximated arbitrarily well by a finite mixture of normal densities with common variance (or covariance matrix in the multivariate case). Thus, mixture models provide a convenient semiparametric framework in which to model unknown distributional shapes, whatever the objective. A mixture model is able to model quite complex distributions through an appropriate choice of its components to represent situations where a single parametric family is unable to provide a satisfactory model for local variations in the observed data. Inferences about the modeled phenomenon can be made

without difficulties from the mixture components. A full account of the theory and applications of mixtures can be found in Karlis & Xekalaki (2003).

3.2 Basic definitions

Let $\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_n$ denote a random sample of size n, where $\boldsymbol{Y}_j$ is a p-dimensional random vector with probability density function $f(\boldsymbol{y}_j)$ on $\mathbb{R}^p$. In practice, $\boldsymbol{Y}_j$ contains the random variables corresponding to p measurements made on the j-th recording of some features on the phenomenon under study. We let $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, \boldsymbol{Y}_2^T, ..., \boldsymbol{Y}_n^T)^T$, where the superscript T denotes vector transpose. Note that we are using $\boldsymbol{Y}$ to represent the entire sample; that is, $\boldsymbol{Y}$ is an n-tuple of points in $\mathbb{R}^p$. Where possible, a realization of a random vector is denoted by the corresponding lower-case letter. For example, $\boldsymbol{y} = (\boldsymbol{y}_1^T, \boldsymbol{y}_2^T, ..., \boldsymbol{y}_n^T)^T$ denotes an observed random sample where $\boldsymbol{y}_j$ is the observed value of the random vector $\boldsymbol{Y}_j$.

Although we are taking the feature vector $\boldsymbol{Y}_j$ to be a continuous random vector here, we can still view $f(\boldsymbol{y}_j)$ as a density in the case where $\boldsymbol{Y}_j$ is discrete by the adoption of counting measure. We suppose that the density $f(\boldsymbol{y}_j)$ of $\boldsymbol{Y}_j$ can be written in the form

$$f(\boldsymbol{y}_j) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{y}_j) \tag{3.2.1}$$

where the $f_i(\boldsymbol{y}_j)$ are densities and π_i are non-negative quantities that sum to one; that is

$$0 < \pi_i \leq 1 \quad (i = 1,2,...,g) \tag{3.2.2}$$

and

$$\sum_{i=1}^{g} \pi_i = 1 \tag{3.2.3}$$

The quantities $\pi_1, \pi_2, ..., \pi_g$ are called the mixing proportions or weights. As the functions $f_1(\boldsymbol{y}_j), f_2(\boldsymbol{y}_j), ..., f_g(\boldsymbol{y}_j)$ are densities, it is obvious that (3.2.1)

defines a density. The $f_i(\boldsymbol{y}_j)$ are called the component densities of the mixture. We shall refer to the density (3.2.1) as a *g*-component finite mixture density and refer to its corresponding distribution $F(\boldsymbol{y}_j)$ as a *g*-component finite mixture distribution.

In this formulation of the mixture model, the number of components *g* is considered fixed. However, in many applications, the value of *g* is unknown and has to be inferred from the available data, along with the parameters in the specified forms for the component densities.

When the number of components is allowed to increase with the sample size *n*, the model is called Gaussian mixture sieve; see Geman and Hwang (1982), Roeder (1992), Priebe and Marchette (1993), Priebe (1994), and Roeder and Wasserman (1997).

3.3 Interpretation of mixture models

An obvious way of generating a random vector $\boldsymbol{Y}_j$ with the *g*-components mixture density $f(\boldsymbol{y}_j)$, given by (3.2.1), is as follows. Let Z_j be a categorical random variable taking on the values 1,2,…,*g* with probabilities $\pi_1,\pi_2,...,\pi_g$, respectively, and suppose that the conditional density of $\boldsymbol{Y}_j$ given $Z_j = i$ is $f_i(\boldsymbol{y}_j)$ (*i*=1,2,…,*g*). Then the unconditional density of $\boldsymbol{Y}_j$ (that is, its marginal density) is given by $f(\boldsymbol{y}_j)$. In this context, the variable Z_j can be thought of as the component label of the feature vector $\boldsymbol{Y}_j$. In later work, it is convenient to work with a *g*-dimensional component label vector $\boldsymbol{Z}_j$ in place of the single categorical variable Z_j, where the *i*-th element of $\boldsymbol{Z}_j$, $Z_{ij} = (\boldsymbol{Z}_j)_i$, is defined to be one or zero, according to whether the component of origin of $\boldsymbol{Y}_j$ in the mixture is equal to *i* or not (*i*=1,2,…,*g*). Thus, $\boldsymbol{Z}_j$ is distributed according to a multinomial distribution consisting of one draw on *g* categories with probabilities $\pi_1,\pi_2,...,\pi_g$; that is,

$$pr\{\boldsymbol{Z}_j = \boldsymbol{z}_j\} = \pi_1^{z_{1j}} \pi_2^{z_{2j}} \cdots \pi_g^{z_{gj}} \tag{3.3.1}$$

We write

$$\boldsymbol{Z}_j = \text{Mult}_g(1, \boldsymbol{\pi}) \tag{3.3.2}$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_g)^T$.

In the interpretation above of a mixture model, an obvious situation where the *g*-component mixture model (3.2.1) is directly applicable is where $\boldsymbol{Y}_j$ is drawn from a population *G* which consists of *g* groups, $G_1, G_2, ..., G_g$ in proportions $\pi_1, \pi_2, ..., \pi_g$. If the density of $\boldsymbol{Y}_j$ in group G_i is given by $f_i(\boldsymbol{y}_j)$ for *i*=1,2,…,*g*, then the density of $\boldsymbol{Y}_j$ has the *g*-component mixture form (3.2.1). In this situation, the *g* components of the mixture can be physically identified with the *g* externally existing groups $G_1, G_2, ..., G_g$.

In biometric applications for instance, a source of heterogeneity is often age, sex, species, geographical origin, and cohort status. In these cases, the population is a mixture of *g* distinct groups that are known a priori to exist in some physical sense.

However, there are also many cases involving the use of mixture models where the components cannot be identified with externally existing groups as above. In some instances, the components are introduced into the mixture model to allow for greater flexibility in modelling a heterogeneous population that is apparently unable to be modelled by a single component distribution. At the extreme end of that, we obtain the nonparametric kernel estimate of a density if we fit a mixture of *g*=*n* components in equal proportions $1/n$, where *n* is the size of the observed sample. For example, if $y_1, y_2, ..., y_n$ denotes an observed (univariate) sample of size *n*, then we obtain the kernel estimate of the density Y_j given by

$$\hat{f}(y_j) = \frac{1}{nh} \sum_{i=1}^{n} k((y_j - y_i)/h) \tag{3.3.3}$$

if in (3.2.1) we set *g*=*n* and $\pi_i = 1/n$ and take $f_i(y_j) = h^{-1}k((y_j - y_i)/h)$ for some kernel function $k(\cdot)$ and parameter *h*. Usually, the kernel $k(\cdot)$, which is a density, has its mode at the origin; see for example, the monographs of Devroye and Györfi (1985), Silverman (1986), and Scott (1992) on nonparametric density estimation.

Thus, for values of the number of components g between 1 and the sample size n, mixture models can be viewed as a semiparametric compromise between (a) the fully parametric model as represented by single (g=1) parametric family and (b) a nonparametric model as represented in the case of g=n by the kernel method of density estimation.

Thus, it can be seen that the mixture models occupy an interesting niche between parametric and nonparametric approaches to statistical estimation. As explained by Jordan and Xu (1996), mixture model based approaches are parametric in that parametric forms $f_i(y_j;\boldsymbol{\theta}_i)$ are specified for the component density functions, but that they can also be regarded as nonparametric by allowing the number of components g to grow. Hence, mixture models have much of the flexibility of nonparametric approaches, while retaining some of the advantages of parametric approaches, such as keeping the dimension of the parameter space down to a reasonable size. Mixture models, therefore, provide a convenient method of density estimation that lies somewhere between parametric models and kernel density estimators.

Concerning the modeling of count data, the fitting of a single Poisson distribution often forces too much structure on the data leading to problem such as overdispersion. The use of a mixture model allows a compromise between the homogeneous Poisson model and nonparametric models, which, although avoiding strong distributional assumptions, have other disadvantages including high-data dependency of model estimates (Böhning et al. 1994; Böhning, 1999).

3.4 Conventional approaches of classification

3.4.1 Mapping percentiles

Let us assume that a certain characteristic of interest is recorded for some aggregated unit such as an area (county, municipality, etc.). Let $x_1, x_2, \ldots, x_n$ be the sample for the n areas under consideration. Then the classification of each area is based on the percentiles of the empirical

distribution of the $x_1, x_2, \ldots, x_n$. Recall that the p-th percentile of some continuous random variable X with distribution function F is given as that value x_p in the sample space of X, for which $F(x_p)=p$, $0<p<1$. This means that each area is classified according to the characteristic's value into the associated percentile.

A map of the areas under consideration is then used to print the classification, using a different color for each obtained group. The researcher can choose which percentiles to use. For example, the median (for p=0.5), the quartiles (for p=0.25, p=0.5, p=0.75) or the quintiles (for p=0.2, p=0.4, p=0.6, p=0.8) can be used. It is obvious that the choice of percentile will force a certain pattern or structure in the map. Therefore, this technique has its deficiencies.

3.4.2 Mapping p-values

This conventional method is often used to construct disease atlases and it based on the standardized mortality ratio (SMR). The SMR is defined as the ratio of observed O and expected E mortality cases, where the number of expected cases is computed on the basis of an external inference population. For the analysis it is assumed that in area i the observed number of deaths O_i follows a Poisson with parameter λE_i:

$$f(o_i, \lambda E_i) = Poisson(o_i, \lambda E_i) = \exp(-\lambda E_i)(\lambda E_i)^{o_i} / o_i!$$

Here, $x_i = o_i / E_i$ is the observed SMR in area i, i=1,2,…n, whereas λ is the theoretical SMR. The conventional display map is based on the p-value under the homogeneous Poisson distribution:

$$P(O_i \geq o_i) = Poisson(o_i, \lambda E_i) + Poisson(o_i + 1, \lambda E_i) + \ldots, \text{ if } x_i \geq \lambda$$

$$P(O_i < o_i) = Poisson(o_i - 1, \lambda E_i) + Poisson(o_i - 2, \lambda E_i) + \ldots + Poisson(0, \lambda E_i), \text{ if } x_i < \lambda.$$

λ is either set to 1 (no increased risk) or replaced by the MLE under homogeneity $\hat{\lambda} = \sum_{i=1}^{n} O_i \Big/ \sum_{i=1}^{n} E_i$.

This method has also a number of deficiencies. The significance of some SMR will depend much on the size of the area, in the sense that areas with small population sizes have greater chances to show significant result. In addition, the method is unable to detect a homogeneous population.

3.4.3 Mapping Empirical Bayes estimates

One of the disadvantages of the previous conventional methods is that they assume a homogeneous population. Practice has shown that it is better to assume that there is a distribution of λ valid in the population. This leads to empirical Bayes estimates. We consider the Bayes risk

$$\int_{\lambda}\int_{x} (\hat{\lambda}(x) - \lambda)^2 f(x/\lambda) p(\lambda) dx d\lambda \qquad (3.4.1)$$

with respect to the Euclidean loss function $(\hat{\lambda} - \lambda)^2$. Here $p(\lambda)$ denotes the distribution of λ in the population, which may be continuous or discrete. If x is a standardized mortality ratio $x=O/E$, for which O conditional on λ is Poisson with mean λE, then E(X)= λ and Var(X)= λ /E.

We are interested in finding a Bayes estimate $\hat{\lambda}(x)$ which minimizes (3.4.1). This can be easily accomplished by considering linear Bayes estimators $\hat{\lambda}(x) = \alpha + \beta x$. In this case the best linear Bayes estimator (Böhning, 1999) $\hat{\lambda}(x) = \alpha + \beta x$ is given by $\beta = \dfrac{\tau^2}{\tau^2 + \mu/E}$ and $a = (1-\beta)\mu$, where μ is the mean and τ^2 is the variance with respect to $p(\lambda)$.

The estimation of μ and τ^2 is the next step. The characterization *empirical Bayes estimates* is a result of this aspect. Conventionally, the marginal density

$$f(o_i, E_i, \Phi) = \int_0^{\infty} Poisson(o_i, \lambda E_i) p(\lambda) d\lambda$$

is considered and maximum likelihood estimates are found with respect to this mixture density. At this point one can either assume that $p(\lambda)$ is parametric, continuous density or leave $p(\lambda)$ be completely nonparametric. In the first case Φ denotes the associated to $p(\lambda)$ parameters, whereas in the second case Φ would coincide with the nonparametric mixing distribution. When $p(\lambda)$ takes specific forms then specific marginal distributions are also achieved. For example, when $p(\lambda; \theta, g)$ is a Gamma distribution we obtain a negative binomial distribution. Therefore, estimating μ and τ^2 leads to maximum likelihood estimation of the parameters of a negative binomial distribution.

The posterior distribution of λ is defined as

$$f(\lambda \mid x) = \frac{f(x \mid \lambda) p(\lambda)}{f(x; P)} = \frac{f(x \mid \lambda) p(\lambda)}{\int_0^{+\infty} f(x \mid \lambda) p(\lambda) d\lambda}$$

if P is a parametric distribution with density $p(\lambda)$, or

$$f(\lambda_j \mid x) = \frac{f(x \mid \lambda_j) p_j}{f(x; P)} = \frac{f(x \mid \lambda_j) p_j}{\sum_{l=1}^{k} f(x \mid \lambda_l) p_l}$$

if P is nonparametric and discrete. Having $f(\lambda \mid x)$ or its estimate available the posterior mean, which we called the empirical Bayes estimate of λ is given as

$$x^{\mathrm{EB}} = E(\lambda \mid x) = \int_{-\infty}^{+\infty} \lambda f(\lambda \mid x) d\lambda .$$

Application of mapping using these empirical Bayes estimates in real data (Böhning, 1999) has shown that this method improves some of the deficiencies of the two conventional ones, yet some problems still remain open.

3.5 The number of subpopulations *g* is unknown

As we noticed before, the number of support points g is not always known a priori and it must also be estimated from the data. In this case we use semiparametric maximum likelihood methods for mixtures by maximizing the likelihood over all the mixing distributions with finite support.

Semiparametric maximum likelihood methods are rather important. As Laird (1978) has shown, if the true mixing distribution is continuous we are restricted to estimate the mixing distribution by a finite-step distribution, i.e. by reducing the mixture model to a finite mixture model with unknown number of support points. Moreover, the number of support points is crucial in many applications since it determines the number of subpopulations comprising the entire population. It is obvious that this case is much more complicated than the case of known g and special algorithms and numerical methods are needed.

We first have to examine under which conditions a semiparametric maximum likelihood (SML) estimate for finite mixtures exists. Lindsay's theorem (1983a) gives sufficient conditions for examining if the global maximum has been obtained.

In order to describe these conditions we need to introduce some notation. Let $D(G,P)$ be the directional derivative of the log-likelihood from the mixing distribution P at the direction of another mixing distribution G.

$$D(G,P) = \lim_{e \to 0} \left[\frac{\ell\left((1-e)P + eG\right) - \ell(P)}{e} \right] \tag{3.5.1}$$

This quantity measures the infinitesimal change of the likelihood when a new distribution G is added to the mixing distribution P. Of special interest is the case when the new mixing distribution G is a degenerate distribution at the point θ. Then we can define the gradient function

$$D(\theta,P) = \sum_{i=1}^{n} \left\{ \frac{f(x_i \mid \theta)}{f(x;P)} - 1 \right\} \tag{3.5.2}$$

$D(\theta,P)$ plays an important role in the case of semiparametric maximum likelihood method for mixtures. Using the gradient function we can state the

following theorem of Lindsay (1983a), which provides sufficient conditions for an estimate to be a SML estimate.

Theorem 3.1 (Lindsay 1983a): $\hat{G}$ is the SML estimate of the mixing distribution G if and only if the following relations hold.

a) $D(\theta,\hat{G}) = 0$, for each θ which is a support point of $\hat{G}$

b) $D(\theta,\hat{G}) \leq 0$, for all other values of θ, not in the support of $\hat{G}$.

This theorem applies also in the case of a multidimensional vector of parameters θ, such as the case of finite normal mixtures with unequal variances.

Theorem 3.1 suggests that all the support points of the SML estimate are the local maxima of the gradient function. It also provides useful tools for calculating the SML estimate.

A natural approach is to apply the EM algorithm for successive values of g. For each value of g we check if the conditions are satisfied, otherwise we proceed with the next value of g. It is interesting that the likelihood can be maximized with few support points and the addition of a further support point will not increase the likelihood. This is the case when the likelihood function for fixed g, has multiple maxima, making the maximum likelihood estimate inconsistent. Pfanzagl (1988), discussed the consistency of the maximum likelihood estimates for mixture models.

Böhning et al. (1994), proposed to check for the conditions of theorem 3.1 by choosing a large number of values in a reasonable interval and checking if the maximum of the gradient function is a value very close to 0 that occurs on the support points of the SML estimate. This lies on the fact that small perturbations due to the computer accuracy may not allow the researcher to calculate a value that is exactly 0. However, such an approach for checking if SML estimate has been found is time demanding and simpler conditions are needed.

The uniqueness of the SML estimator has been showed by Simar (1976) for the case of Poisson mixtures. Lindsay (1983a, b) and Lindsay and Roeder (1992, 1993) showed that the SML estimator is unique for members of the

continuous exponential family. They also showed that the SML estimator is unique for discrete mixtures if and only if the probability distribution evaluated using this SML estimator of the mixing distribution coincides with the observed relative frequency distribution. Obviously, for discrete distributions with support in the positive axis, like the Poisson distribution, the probability function evaluated using the SML estimate of the mixing distribution will give positive probability to values greater than the maximum observed value. Hence, this estimated probability distribution will not coincide with the observed relative frequency distribution.

The results of theorem 3.1 provide useful guides for checking if the global maximum is obtained. If we plot the gradient function, this ought to have 0 values at all points on the support of the solution, and if the global maximum is attained, is ought to be restricted down the 0 line. Thus, by plotting the gradient function we can check if the solution is the SML solution. Otherwise, if there exist points outside the support of the solution, for which the gradient is 0, or points with a positive gradient, we have to add points because the global maximum has not been attained.

3.6 The number of support points

The maximum likelihood estimate for a *g*-finite mixture is not necessarily the SML estimate. It is just the best possible solution with the given number of support points. The natural question at this point is whether we know something about the number of support points. Simar (1976), was the first one who concentrated on the particular case of maximum likelihood estimation for Poisson mixtures. He provided the following theorem concerning the number *g* of support points.

Theorem 3.2 (Simar, 1976). If *g* denotes the number of support points of the SML estimate of the mixing distribution, and *N* represents the largest observed value then

a) If $\lambda_1 = 0$ then $k \leq \left[\frac{N+2}{2}\right]$, while if $\lambda_1 > 0$ then $k \leq \left[\frac{N+1}{2}\right]$

b) In every case $k \leq q$

where [a] is the integer part of a, and q is the number of distinct values in the sample.

Laird (1978), conjectured that the number of support points in mixtures from continuous densities cannot be larger than the sample size. She also gave an interesting guide to this search. For a mixed distribution the problem of counting the number of support points is equivalent to counting the number of modes of a mixture of n conjugate densities. For example, for the case of the Poisson probability function, the Gamma density is the conjugate. So if we have assumed a mixed Poisson probability function, then we take a mixture of n Gamma distributions, with parameters $x_i + 1$ and 1 respectively, $i = 1,2,...,n$, and we count the number of modes. This approach gives us an upper bound for the number of support points.

Lindsay (1983a), proved the conjecture of Laird (1978), namely that the number of components cannot be larger than the sample size. Lindsay and Roeder (1993), gave a result similar to Simar's for general discrete distributions.

Intuitively, when we try to estimate a mixed Poisson distribution with g support points the number of estimated parameters is $2g$-1. If we have observed only N different values, then with N parameters we can theoretically fully reconstruct the observations (since we then have a non-linear system of N equations with N unknowns). Therefore, we need to restrict the number of support points. Adding one more component implies that the unknown values to be estimated exceed the number of estimating equations leading to intractabilities. Since we are only interesting in maximizing the likelihood and not in solving the system of equations explicitly, the problem lies in that the constraints for the maximization are too many.

However, when the mixture is discrete, the restriction of the number of support points prevents us from estimating the continuous mixing distribution with a finite approximation with many support points and hence closer to a continuous one. A simple example is the case of a Gamma mixing distribution for a mixture of the Poisson distribution, which leads to the negative binomial

distribution. If the mean is not large, the estimation will provide us with an estimate with few support points which will not resemble the true Gamma mixing distribution.

3.7 Algorithms for Semiparametric Maximum Likelihood estimation for mixtures

In this section we will present some of the algorithms used in order to obtain the semiparametric maximum likelihood (SML) estimate of the mixing distribution. As we noticed before, the number of support points g is unknown, making the procedure of the estimation complex.

Although a simple manner to deal with this problem would be to derive the maximum likelihood estimate for successive values of g using the EM algorithm and the conditions of theorem 3.1 as a stopping rule, this would require a lot of computational effort. More sophisticated algorithms have been proposed in literature, using special methods of numerical analysis. The main idea for them is to start with an initial solution with a few support points, usually one or two, and then add one or more new points at each step, sometimes replacing old but "bad" ones until some criteria are fulfilled. In the following paragraphs a description of these algorithms is given.

For the properties of the semiparametric maximum likelihood estimate of the mixing distribution see Karlis (2001).

3.7.1 The Vertex Direction Method (VDM)

The Vertex Direction Method (VDM) uses the gradient function defined in (4.5.2) in order to obtain the SML estimate of the mixing distribution. As in most methods, we start with some initial value P^0 for the mixing distribution. P^i, in general, represents the estimate of the mixing distribution after i steps. In each step we add as a new point the value of θ which maximizes the gradient function. The probability associated with this new

support point must be calculated so that for the new estimate P^{i+1} the log-likelihood increases, namely $\ell(P^{i+1}) \geq \ell(P^i)$. Generally, P^{i+1} is a convex combination of P^i and P_θ, where P_θ is a distribution which puts all its mass at the point θ, thus it is a degenerate distribution. In other words, $P^{i+1} = (1-\alpha)P^i + \alpha P_\theta$, where α is obviously the probability assigned at the new support point.

Therefore, the VDM algorithm consists of the following steps:

Step 1: Find $\theta_{\max}$ to maximize $D(\theta, P^i)$.

We are therefore interested in finding a new "good" support point. By maximizing $D(\theta, P^i)$ we find the best point to the direction towards the new estimate P^{i+1}. Thus, the new support point is $\theta_{\max}$.

Step 2: Find α to ensure that $\ell\left((1-\alpha)P^i + \alpha P_{\theta_{\max}}\right) \geq \ell(P^i)$.

At this step we construct the new estimate, adjusting the probabilities of the "old" support points so that the new probability estimates add up to 1.

Step 3: Examine if a global maximum is attained by means of the conditions of theorem 4.1, otherwise go back to step 1.

It is obvious that steps 1 and 2 require a lot of numerical work. At step 1, maximization can be achieved by initially searching in a grid of distinct points in some interval. A good choice for such an interval is between 0 and the maximum observed value in the sample. Then one may start from the point where the gradient function has its maximum to locate the maximum by some iterative scheme like the Newton-Raphson. The strategy is that the grid search reaches the maximum, which is then located easily via a standard maximization algorithm. The process is carried out by searching for the point where the derivative is 0. However, this may lead to the minimum near the point instead of the maximum itself. Thus, we have to check if the obtained value is a minimum or a maximum.

The main problem associated with this step, is that the maximum may lie outside the admissible range. It may be negative or the gradient function may

increase to the infinity. In both circumstances the maximum cannot be found and the algorithm stops.

In step 2 the value of α must be found. Böhning (1989, 1995), describes algorithms for finding a value for α. He calls these algorithms as monotone step algorithms. He shows that the problem of finding the value of α, can be reduced to a problem of estimating a closed area. Hence, algorithms used for estimating an area are useful for finding a value for α.

Apart from the monotone step algorithms another choice would be to find an α which maximizes $\ell(P^{i+1})$ with respect to α. Böhning (1995) shows that $\ell(P^{i+1})$ is concave with respect to α and thus a maximum value exists which is very easy to locate by a numerical algorithm.

In case we have a restriction for the number of support points we must add this condition at step 3. This holds for the Poisson case.

The algorithm itself has some serious disadvantages. The first is that it is very slow in its convergence behavior. Böhning (1995), proposed some improvements for the VDM. These improvements, however, had a marginal effect because of some inherent disadvantages. In addition, the algorithm is very sensitive to the initial value (or values). Inappropriate initial values can destroy the algorithm. This is due to the fact that in step 1 we are not able to find a maximum since the quantity $D(\theta, P^i)$ is monotonic or because the maximum is outside the admissible range. On the other hand, the initial point, even if it is a "bad" one, remains at the final estimated mixing distribution, since the algorithm only adds points without removing the old "bad" ones. This may cause the destruction of the algorithm. For the Poisson case where the number of support points is usually small, the algorithm is not satisfactory since we have to estimate very few support points. This is not a sever problem in cases where there is no such a strict limitation for the number of support points, as in the case of mixtures of normal or other continuous distributions, since the effect of the initial point is negligible after the adding of a large number of new points.

A further problem is that the mean of the estimated mixing distribution should be equal to the one estimated from the sample. This is due to the fact that the gradient function and its derivative evaluated at the support points

ought to be zero in order for the estimate to be a maximum likelihood estimate. This complicates the steps of the VDM algorithm.

For further discussion on this algorithm see Karlis (2001).

3.7.2 The Vertex Exchange Method (VEM)

The Vertex Exchange Method (VEM) tries to overcome some of the advantages of the VDM algorithm. It is faster in convergence than the VDM algorithm. It also deals with the problem of "bad" support points. The main idea of this method is to exchange "good" vertex directions for "bad" ones that are already in support of the current mixing distribution. To do that, there is an additional step in which it finds the worst of the old support points and examines if the new point $\theta_{\max}$ can replace at all the "bad" point $\theta_{\min}$. In this case the "bad" point is eliminated. Lesperance and Kaldbfeish (1992), described the algorithm in detail.

The algorithm consists of the following steps:

Step 1: Find $\theta_{\max}$ to maximize $D(\theta, P^i)$ over all possible values of θ.

This step leads to the new support point. Grid search with a complementary numerical search is a useful tool for finding it.

Step 2: For all the points in the support of P^i calculate the gradient function $D(\theta, P^i)$. Find the point $\theta_{\min}$ which has the minimum value over all the support points.

Step 3: Set $P^{i+1} = P^i + \alpha p^* (P_{\theta_{\max}} - P_{\theta_{\min}})$, where p^* is the probability of the "bad" support point.

The meaning of this expression is that we take some proportion α of the probability of the "bad" support point and we assign it to the new support point. If α=1, we reject the "bad" support point. If α=0 we do not change our estimate at all. The problem is again to find the value of α to ensure that $\ell(P^{i+1}) \geq \ell(P^i)$. As before, a monotone step-length algorithm or direct maximization are possible methods.

Step 4: Examine if the global maximum is obtained.

With the VEM method it is not necessary to add a point at every iteration. However, if the new point is "bad" then the algorithm will fail. This method is more dynamic than VDM and it converges quicker than the VDM algorithm. Lesperance and Kaldbfeish (1992), gave examples to show the superiority of the VEM algorithm. Böhning (1995), suggested some slight improvements of the algorithm.

The choice of initial values is still a problem. "Bad" initial values may lead to "bad" choices of new points and, since the algorithm exchanges one point with another at each iteration, it may delay very much to get rid of the "bad" points. For the Poisson case, where the number of admissible support points is small, the VEM algorithm does not give satisfactory results.

3.7.3 The Intra Simplex Direction Method (ISDM)

The VDM and VEM algorithms and their modifications have the computational disadvantage that one must keep track of and perform computations over the complete accumulated set of support points at each iteration. All of them add at the most one new support point.

Lesperance and Kaldbfeish (1992), proposed another method, known as the Intra-Simplex Direction Method (ISDM). In contrast to the other two methods, at each step of ISDM several new points are found instead of one. At step 1, instead of finding the global maximum, several local maxima points $\theta_1^*, \theta_2^*, ..., \theta_r^*$ are found. Then we must find the probability assigned to each of these points maximizing the corresponding likelihood as in step 2 of the VDM and VEM algorithms. It is obvious that this method requires a lot of computational work. In general, it is hard to obtain all the local maxima and we need a very careful search to do so. In fact, the added labour is in step 2, because from the grid search for the maximum of step 1 we have already calculated the gradient function for several values of θ. The EM algorithm is an adequate choice for finding the probabilities of step 2. As Lesperance and

Kaldbfeish (1992) point out the complicated computations at each iteration are compensated by the smaller number of iterations until the global maximum is attained. Again, this approach is not appropriate for the Poisson case since, as we said before, the number of iterations is usually small.

3.7.4 Related algorithms

Dersimonian (1986, 1990), proposed an algorithm similar to Simar's that uses the conditions given by Lindsay (1983a) for examining if the maximum is obtained. Her algorithm treated the case of mixtures of normal, exponential, binomial and Poisson distributions, starting from a uniform estimate. She assigned equal probabilities to equally spaced points. It uses the EM algorithm until some kind of convergence is achieved and then, by maximizing the gradient function, it finds a new support point. Then the EM algorithm is applied so as to maximize the likelihood for the new set of support points. The algorithm stops when we cannot add a new support point or the conditions of Lindsay (1983a) are satisfied.

An interesting connection with algorithms used in the field of D-optimal designs is discussed by Böhning (1989, 1995). He showed that searching for the ML estimate of a mixing distribution is equivalent to searching for a D-optimal design. Hence, results from this field are applicable. This is the reason why, in some cases, algorithms appear with different names. The author cited a large number of references. In the same paper he described in detail the monotone step-length algorithms. Lindsay (1983a, b), also described the similarity with D-optimal designs.

Heckman and Singer (1984), used the SML estimate for estimating the mixing distribution in duration models. They showed how much sensitive is the estimation based on a specified mixing distribution and they proposed the SML estimate as a method to avoid biasing the results by choosing an arbitrary mixing distribution. Böhning (1989), described the geometry of the likelihood of mixtures. Similar is the work of Lindsay (1983a, b). They both showed pictorially, in a few dimensions, how we proceed to maximize the likelihood, giving an excellent insight to the whole procedure. Also this approach gives some knowledge about how we can improve our search.

Lindsay (1995) gave some bounds for the possible improvement of the likelihood in each step of the VDM algorithm.

Constrained maximization is described by Lesperance and Kaldbfeish (1992), Böhning (1995) and Susko et al. (1997). Lesperance and Kaldbfeish (1992) proposed a semi-infinite programming routine for the maximization, which seems to work very well.