EM ALGORITHM FOR MIXED POISSON AND OTHER DISCRETE DISTRIBUTIONS

BY

DIMITRIS KARLIS

ABSTRACT

Mixed Poisson distributions are widely used in various disciplines including actuarial applications. The family of mixed Poisson distributions contains several members according to the choice of the mixing distribution for the parameter of the Poisson distribution. Very few of them have been studied in depth, mainly because of algebraic intractability. In this paper we will describe an EM type algorithm for maximum likelihood estimation for mixed Poisson distributions. The main achievement is that it reduces the problem of estimation to one of estimation of the mixing distribution which is usually easier. Variants of the algorithm work even when the probability function of the mixed distribution is not known explicitly but we have only an approximation of it. Other discrete distributions are treated as well.

KEYWORDS

Poisson mixtures, beta-binomial distribution, Yule distribution, Monte Carlo EM.

1. INTRODUCTION

Starting from a distribution family $f(x|\theta)$ we may obtain a very rich new family of distributions if we allow the parameter θ to be itself a random variable with distribution function $G(\theta|\varphi)$ depending on a vector of parameters φ. Note that θ is not necessarily a scalar and it can be vector-valued. Then the unconditional distribution of x will be given by

$$f(x|\theta) = \int f(x|\theta)\, dG(\theta|\varphi) \tag{1}$$

Usually, $f(x|\varphi)$ is called the mixed distribution while $G(\theta|\varphi)$ is called the mixing distribution. Mixture models are also called overdispersion models, because keeping the mean fixed they have variances larger than the original model. For a thorough treatment of mixture models the reader can refer to Titterington et al (1985), Lindsay (1995), Bohning (1999), McLachlan and Peel (2000).

ASTIN BULLETIN, Vol. 35, No. 1, 2005, pp. 3-24

Mixed Poisson distributions are widely used in actuarial problems to model the claim process (see, e.g. Grandell, 1997). The family of mixed Poisson contains a large number of members. Only few of them have been used in practice mainly due to algebraic problems appearing when attempting to use them in real data situations. The purpose of this paper is to illustrate that Maximum Likelihood (ML) estimation can be accomplished rather easily via EM type algorithms that use the inherent latent structure of mixture models. The main focus of the paper will be put on mixed Poisson distributions, however the case of other families will also be discussed. The contribution of the present paper lies mainly on the specific application of the general EM algorithm for mixtures in the mixed Poisson setting. We provide helpful devices and a general framework to handle this family of distributions. However, the algorithms derived for certain distributions facilitate ML estimation and, thus, enhances the applicability of them in real data problems.

The remaining of the paper proceeds as follows. Section 2 provides some background material for the EM algorithm. In section 3, the general theory related to mixed Poisson distributions is described, while the algorithm is applied to a wide variety of mixed Poisson distributions in section 4. A real data application can be found in section 5. A Monte Carlo EM algorithm is described in section 6. Extensions to other discrete distributions that can be seen as arising by mixtures are described in section 7. Concluding remarks can be found in section 8.

2. The EM Algorithm for Mixtures

The EM algorithm (Dempster *et al.*, 1977) is a powerful algorithm for ML estimation for data containing missing values or being considered as containing missing values. This formulation is particularly suitable for distributions arising as mixtures since the mixing operation can be considered as producing missing data. An important feature of the EM algorithm is that it is not merely a numerical technique but it also offers useful statistical insight.

Suppose that the complete data $Y_i = (X_i, Z_i)$ consist of an observable part X_i and an unobservable part Z_i. When the direct maximization of $\log p(X|\varphi)$ with respect to the parameter φ is not easy, the algorithm augments the observed data to a set of complete data which can be reduced to the observed data via a many to one mapping. The EM algorithm maximizes $\log p(X|\varphi)$ by iteratively maximizing $E(\log p(Y|\varphi))$. At the E-step of the $(k+1)-th$ iteration the expected loglikelihood of the complete data model is calculated as $Q(\varphi|\varphi^{(k)}) = E(\log p(Y|\varphi)|\ X, \varphi^{(k)})$ where the expectation is taken with respect to the conditional distribution $f(Y|X, \varphi^{(k)})$ and then, at the M-step, $Q(\varphi|\varphi^{(k)})$ is maximized over φ. When the complete model is from the exponential family then the E-step computes the conditional expectations of its sufficient statistics. This is quite helpful in our case.

Let us return to the mixture formulation. The unobserved quantities are simply the realizations θ_i of the unobserved mixing parameter for each data point X_i. Hence at the E-step one needs to calculate the conditional expectation of some functions of θ_i's and then to maximize the likelihood of the complete model which reduces to maximizing the likelihood of the mixing density. In the case of mixtures from the exponential family the conditional expectations coincide with the sufficient statistics needed for ML estimation of the mixing distribution. Formally, the algorithm can be described as:

- E-Step – Using the current estimates $\varphi^{(k)}$ taken from the $k-th$ iteration, calculate the pseudovalues $t_{ij} = E(h_j(\theta) \mid X_i, \varphi^{(k)})$, for $i = 1, \ldots, n$, $j = 1, \ldots, m$, where $h_j(.)$ are certain functions.
- M-Step – Use the pseudovalues t_{ij} from the E-step to maximize the likelihood of the mixing distribution and obtain the updated estimates $\varphi^{(k+1)}$.
- If some terminating condition is satisfied then stop iterating otherwise go back to the E-step for more iterations.

The M-step is somewhat obvious and depends on the assumed mixing distribution. For some distributions a special iterative scheme may be appropriate, and perhaps another EM algorithm. The E-step however is not straightforward. For linear functions of θ the conditional posterior expectations can be easily and accurately obtained as it will be shown in the next section. For more complicated functions, if exact solution is not available, one may proceed either by appropriate approximations based on Taylor approximations or by numerical approximations including numerical integration and/or simulation based approximations. All the above solutions seem to work well in practice.

All the controversies for and against the EM algorithm apply. The convergence is usually slow, but this depends on the unobserved information that needs to be estimated at the E-step. Usually no singularities on the likelihood surface are encountered and thus good initial values are useful only to speed up the convergence and not to locate the global maximum. It is worth mentioning that the by-products of the algorithm are useful for further inference. For example, in the mixed Poisson case, posterior expectations of the form $E(\theta | x)$ can be used to predict future outcomes or for Empirical Bayes estimation. In actuarial practice experience rating can be based on this quantity as well. Such quantities are calculated during the execution of the algorithm and they are readily available after the convergence of the EM algorithm. Recall that the EM is also useful for maximum a posteriori estimation in the Bayesian setting (see, e.g. Carlin and Louis, 1996) and, thus, the algorithms can be used for Bayesian estimation as well.

3. Poisson Mixtures

Assume $f(x|\theta) = \exp(-\theta)\theta^x / x!$, $x = 0,1,\ldots$. i.e. the Poisson distribution with parameter $\theta > 0$. Then the resulting mixed distribution is a mixed Poisson distribution and its probability function is given by

$$P_g(x|\varphi) = \int_0^\infty \frac{\exp(-\theta)\theta^x}{x!} dG(\theta|\varphi) \tag{2}$$

where the subscript denotes the mixing density. Some of the well known discrete distributions can be obtained as mixed Poisson distributions, like the negative binomial distribution. Note that if $G(\theta|\cdot)$ is a finite step distribution the family of finite Poisson mixtures arises.

In the literature there are several mixed Poisson distributions (see Johnson *et al.*, 1992). A large list of distributions in this family can be seen in Karlis (1998). However, very few of them have been studied in depth, mainly because of lack of algebraic convenience.

As it is shown at the previous section, at the E-step one needs to calculate the posterior expectation of some function $h(\theta)$. We will now see that if this function is of a certain form, exact calculations are easily obtainable for the more general case of mixtures from the power series family of distributions.

Definition: A discrete distribution is said to belong to the power-series family of distributions if its probability function is given by $P(x|\theta) = \alpha_x \theta^x (A(\theta))^{-1}$, $x = 0, 1, \ldots$, with $\alpha_x > 0$ and $A(\theta)$ is a function of θ not depending on x.

Many of the well known discrete distributions belong to this family, like the Poisson, the binomial, the negative binomial and other distributions. Suppose now that the parameter θ is itself a random variable. Then we have a power series mixture with probability function

$$P_g(x|\varphi) = \int_\theta \frac{\alpha_x \theta^x}{A(\theta)} g(\theta|\varphi)\, d\theta. \tag{3}$$

Then the following result holds:

Lemma 1 (*Sapatinas*, 1995) The posterior expectation $E(\theta^r|x)$ where x conditional on θ follows a power series discrete distribution and θ has pdf $g(\theta)$ is given by:

$$E\left(\theta^r \middle| x\right) = \frac{P_g(x+r|\varphi)\alpha_x}{P_g(x|\varphi)\alpha_{x+r}}$$

where $P_g(x|\varphi)$ is the power series mixture defined in (3).

For the Poisson distribution $\alpha_x = 1/x!$ and hence the posterior expectations are given by

$$E\left(\theta^r \middle| x\right) = \frac{(x+r)!\, P_g(x+r|\varphi)}{x!\, P_g(x|\varphi)} \tag{4}$$

with $P_g(x|\varphi)$ given in (2).

Note that we may extend the above results to the case of negative r, when $(x+r) > 0$. This enables one to find for example posterior expectations of the form $E(\theta^{-r}|x)$, $r > 0$. In addition other expectations can be obtained through the above generic formulas by using Taylor expansions. At the next section we will give a variety of results concerning Poisson mixtures.

4. Application to mixed Poisson distributions

In general we assume that for each observation X_i, $i = 1,\ldots,n$ the distribution of $X_i|\theta_i$ is the Poisson distribution, while θ_i varies according to a mixing density. We will denote as θ the mixing variable in general, while θ_i will be used to denote the realization for the i-th observation.

4.1. Geometric Distribution

Suppose that $g(\theta|\lambda) = \lambda\exp(-\theta\lambda)$, θ, $\lambda > 0$, i.e. θ_i follows an exponential distribution with mean λ^{-1}, then the resulting geometric distribution has probability function given by:

$$P(x|\lambda) = \left(\frac{\lambda}{1+\lambda}\right)\left(\frac{1}{1+\lambda}\right)^x, \tag{5}$$

$x = 0, 1,\ldots$, $\lambda > 0$. Using the above formulation the ML for the parameter λ for both the exponential and the geometric distributions is $\hat{\lambda} = \bar{x}^{-1}$. It is clear that an EM scheme is not useful at all, but it constitutes the most simple example. The EM scheme is constructed by updating the current estimate λ_{old} with the new estimate as:

E-step: Calculate the pseudovalues

$$t_i = E(\theta_i|x_i) = \frac{(x_i+1)\,P(x_i+1|\lambda_{\text{old}})}{P(x_i|\lambda_{\text{old}})} = \frac{x_i+1}{\lambda_{\text{old}}+1}$$

for $i = 1,\ldots,n$

M-step: Find the new estimate λ_{new} by:

$$\lambda_{\text{new}} = \frac{n}{\sum_{i=1}^{n} t_i} = \frac{\lambda_{\text{old}}+1}{\bar{x}+1}$$

If some criterion is satisfied stop iterating else go back to E-step.

It is easy to see that the iteration stops when $\lambda = \bar{x}^{-1}$

4.2. Poisson-Lindley distribution

Sankaran (1970) proposed the Poisson-Lindley distribution for the analysis of count data. The distribution arises from the simple Poisson distribution if the parameter θ follows the Lindley distribution having density function

$$g(\theta|p) = \frac{p^2}{p+1}(\theta+1)\exp(-\theta p), \quad \theta, p > 0 \tag{6}$$

The resulting Poisson-Lindley distribution has probability function given by

$$P(x|p) = \frac{p^2(p+2+x)}{(p+1)^{x+3}}, \quad x = 0, 1, \ldots, p > 0$$

Sankaran (1970) didn't give MLE for the parameter p because of the computational difficulty to do so. An EM scheme can be easily derived. The MLE for the parameter p from a sample $X_1, X_2, \ldots, X_n$ from the the Lindley distribution in (6), is given by the solution of the equation $\frac{p+2}{p(p+1)} = \bar{x}$, where $\bar{x}$ is the sample mean. Thus, an EM scheme is as follows:

E-step: Calculate the pseudovalues

$$t_i = E(\theta_i|x_i) = \frac{(x_i+1)P(x_i+1|p_{\text{old}})}{P(x_i|p_{\text{old}})} = \frac{(p_{\text{old}}+x_i+3)(x_i+1)}{(p_{\text{old}}+x_i+2)(p_{\text{old}}+1)}$$

for $i = 1, \ldots, n$. From these values calculate $\bar{t} = \sum_{i=1}^{n} t_i / n$

M-step: Find the new estimate p_{new} by:

$$p_{\text{new}} = \frac{-(\bar{t}-1) + \sqrt{\bar{t}^2 + 6\bar{t} + 1}}{2\bar{t}}$$

If some criterion is satisfied stop iterating else go back to E-step.

Sankaran's moment estimate is the same as the M-step if $\bar{t}$ is replaced by $\bar{x}$. This verifies the conjecture of Sankaran that the moment estimate is close to the MLE.

4.3. Hermite distribution (Poisson-Normal)

Hermite distribution, examined in Kemp and Kemp (1965) can be considered as a generalized (or compound for some authors) Poisson distribution, namely the distribution of the random variable $Y = X_1 + X_2 + \ldots + X_N$, where N follows a Poisson distribution and each X_i, $i = 1, \ldots, N$ follows a Binomial$(2, p)$ distribution. Formally the distribution can also be considered as a mixed Poisson, with a normal mixing distribution (Kemp and Kemp, 1966). This result lacks any physical interpretation since the parameter of the Poisson distribution ought to be positive. However, if the normal mixing distribution has positive mean much larger than its variance then the probability of a negative value is almost negligible. Recall that many continuous distributions tend to the normal distribution when their parameters take specific values. Considering that $g(\theta|\cdot)$ is a normal distribution with mean μ and variance σ^2, then the resulting

Hermite distribution has probability function given by (see, e.g. Johnson *et al.*, 1992)

$$P(x|\alpha_1,\alpha_2) = \exp(-(\alpha_1+\alpha_2))\sum_{j=0}^{[x/2]}\frac{\alpha_1^{x-2j}\alpha_2^{j}}{(x-2j)!\,j!}, \quad x = 0,1,\ldots,\alpha_1,\alpha_2 > 0 \qquad (7)$$

where, $\alpha_1 = \mu - \sigma^2$ and $\alpha_2 = \sigma^2/2$, and $[\alpha]$ is the integer part of α. The probabilities can be easily calculated via the following iterative scheme: $P(0|\alpha_1,\alpha_2) = \exp(-(\alpha_1+\alpha_2))$, $P(1|\alpha_1,\alpha_2) = P(0|\alpha_1,\alpha_2)\alpha_1$ and $(x+1)P(x+1|\alpha_1,\alpha_2) = \alpha_1 P(x|\alpha_1,\alpha_2) + 2\alpha_2 P(x-1|\alpha_1,\alpha_2)$, $x > 1$. It holds that $E(X) = \alpha_1 + 2\alpha_2 = \mu$ and $Var(X) = \alpha_1 + 4\alpha_2 = \mu + \sigma^2$, while $Var(X)/E(X) = 1 + \frac{2\alpha_2}{\alpha_1+2\alpha_2} = 1 + \frac{\sigma^2}{\mu}$ and hence $1 < Var(x)/E(x) < 2$. Moreover, in order for the normal mixture representation to be valid, it must hold that $\mu - 3\sigma > 0$ in order the probability of a normal variate to be negative to be negligible. The above restrictions suffice for many applications. The EM scheme is as follows. From the current estimates μ_{old} and σ_{old}

E-step: Calculate the pseudovalues

$$t_i = E(\theta_i|X_i) \quad \text{and} \quad s_i = E(\theta_i^2|X_i)$$

using (4), for $i = 1,\ldots,n$.

M-step: Find the new estimates μ_{new} and σ^2_{new} by:

$$\mu_{\text{new}} = \frac{\sum_{i=1}^{n} t_i}{n} \quad \text{and} \quad \sigma^2_{\text{new}} = \frac{\sum_{i=1}^{n} s_i}{n} - \left(\frac{\sum_{i=1}^{n} t_i}{n}\right)^2$$

If some criterion is satisfied stop iterating else go back to E-step.

4.4. Poisson-Inverse Gaussian distribution

Let $IG(\gamma,\delta)$ to denote the Inverse Gaussian distribution with parameters γ and δ and probability density function given by

$$g(\theta|\gamma,\delta) = \frac{\delta}{\sqrt{2\pi}}\exp(\delta\gamma)\,\theta^{-3/2}\exp\left(-\frac{1}{2}\left(\frac{\delta^2}{\theta} + \gamma^2\theta\right)\right), \quad \theta,\gamma,\delta > 0.$$

By using the $IG(\gamma,\delta)$ distribution as a mixing density the Poisson-Inverse Gaussian distribution arises. Its probability function is rather complicated, but simple recurrence relations exist for calculation of the probabilities. So, one can calculate the probabilities using the following iterative scheme:

$$P(0|\gamma,\delta) = \exp(\delta(\gamma-\xi)), \qquad P(1|\gamma,\delta) = \frac{\delta P(0|\gamma,\delta)}{\xi} \qquad \text{and}$$

$$P(x|\gamma,\delta) = \frac{1}{x\xi^2}\left((2x-3)\,P(x-1|\gamma,\delta) + \frac{\delta^2}{x-1}P(x-2|\gamma,\delta)\right),\ \ x = 2,3,...$$

where $\xi = (2+\gamma^2)^{1/2}$. The distribution has been examined by many authors (see, e.g. Sichel, 1974, 1982, among others).

The Inverse Gaussian distribution is a special case of the more general family of Generalized Inverse Gaussian (GIG) distributions with density function

$$g(\theta|\lambda,\gamma,\delta) = \left(\frac{\gamma}{\delta}\right)^{\lambda}\frac{\theta^{\lambda-1}}{2K_\lambda(\delta\gamma)}\exp\left(-\frac{1}{2}\left(\frac{\delta^2}{\theta}+\gamma^2\theta\right)\right),$$

where $K_r(\cdot)$ denotes the modified Bessel function of order r. This distribution will be denoted as $GIG(\lambda,\gamma,\delta)$. Details for the GIG distribution can be found in Jorgensen (1982). The Inverse Gaussian distribution arises when $\lambda = -1/2$. The moments around the origin of the GIG distribution are given by

$$E\left(z^r\right) = \left(\frac{\delta}{\gamma}\right)^r\frac{K_{\lambda+r}(\delta\gamma)}{K_\lambda(\delta\gamma)} \qquad (8)$$

for $r = \dots, -1, 0, 1, \dots$. The $GIG(\lambda,\gamma,\delta)$ is conjugate in the Bayesian sense for the Poisson distribution. Thus if $\theta \sim GIG(\lambda,\gamma,\delta)$ then $\theta|x \sim GIG(\lambda+x, \sqrt{2+\gamma^2}, \delta)$.

An EM scheme is as follows. From the current estimates δ_{old} and γ_{old} the new estimates will be obtained as follows:

E-step: Calculate the pseudovalues $t_i = E(\theta_i|X_i)$ and $s_i = E(\theta_i^{-1}|X_i)$ using (4) and/or (8), for $i = 1,\dots,n$. For $x = 0$ we use the fact that $\theta|x \sim GIG(x-1/2, \xi, \delta)$ and hence $E(\theta_i|X_i = 0) = \delta^{-2}(\delta\xi+1)$

M-step: Find the ML estimates for the parameters of an Inverse Gaussian, using the posterior expectations for θ_i and θ_i^{-1}. We can see that the ML estimates for the parameters of the $IG(\gamma,\delta)$ distribution can be found using the quantities $\hat{M} = \sum_{i=1}^{n} s_i/n$ and $\hat{\Lambda} = n\left(\sum_{i=1}^{n}(t_i - \hat{M}^{-1})\right)^{-1}$ and then $\delta_{\text{new}} = \hat{\Lambda}^{1/2}$ and $\gamma_{\text{new}} = \delta_{\text{new}}/\hat{M}$.

If some criterion is satisfied stop iterating else go back to E-step.

Note that standard maximization of the loglikelihood of the Poisson-Inverse Gaussian distribution involves the derivatives of the Modified Bessel function, which is avoided via the EM algorithm.

4.5. Negative binomial distribution

The negative binomial distribution is widely known as a prominent member of the family of mixed Poisson distributions. If $Gamma(\alpha,\beta)$ denotes the gamma distri-bution with parameters α and β, with density given by $g(\theta|\alpha,\beta) = \theta^{\alpha-1}\exp(-\beta\theta)\,\beta^\alpha/\Gamma(\alpha)$, $\alpha,\beta > 0$ and then, allowing θ to vary according to a $Gamma(\alpha,\beta)$ distribution, one obtains the negative binomial distribution with probability function given by

$$P(x|\alpha,\beta) = \frac{\Gamma(x+\alpha)}{x!\,\Gamma(\alpha)}\left(\frac{\beta}{1+\beta}\right)^{\alpha}\left(\frac{1}{1+\beta}\right)^{x} \tag{9}$$

for $x = 0, 1, \ldots, \alpha, \beta > 0$. It was firstly derived as a Poisson mixture by Greenwood and Yule (1920) in their fundamental paper. The mean equals α/β while the variance equals $\alpha/\beta^2 + \alpha/\beta$. If $\beta \to \infty$ then the negative binomial distribution tends to the Poisson distribution.

ML estimation has been proposed by several authors (see, e.g Piegorsch, 1990). Numerical methods are needed for solving the system of the equations involved. Moreover it has been noticed, that the ML estimates do not exist when the sample mean exceeds the sample variance (see, e.g. Wang, 1996). It is known that if $\theta \sim Gamma(\alpha,\beta)$ distribution, the posterior of $\theta|x$ is a $Gamma(\alpha+x, \beta+1)$ distribution. Applying the EM algorithm we need to obtain $E(\theta|x)$ and $E(\log\theta|x)$. Since the posterior densities are gamma, in fact we need these expectations for a gamma variate. Details can be found in the Appendix. Thus, the EM scheme is given as follows.

From the current estimates α_{old} and β_{old} the new estimates will be obtained as

E-step: Calculate the pseudovalues

$$t_i = E(\theta_i|x_i) = \frac{x_i + \alpha_{\text{old}}}{1+\beta_{\text{old}}} \quad \text{and} \quad s_i = \Psi(\alpha_{\text{old}} + x_i) - \log(\beta_{\text{old}} + 1)$$

for $i = 1, \ldots, n$, $\Psi(\cdot)$ denotes the digamma function.

M-step: Using t_i and s_i the likelihood of gamma variates must be maximized. This can be done relatively easily using the ECM algorithm given in Meng and Rubin (1993). So, update α and β as $\beta_{\text{new}} = \alpha_{\text{old}}/\bar{t}$ and

$$\alpha_{\text{new}} = \alpha_{\text{old}} - \frac{\Psi(\alpha_{\text{old}}) + \log(\beta_{new}) - \bar{s}}{\Psi_3(\alpha_{\text{old}})}$$

where $\Psi_3(x)$ denotes the trigamma function defined as $\vartheta\Psi(x)/\vartheta x$, i.e. the derivative of the digamma function. The M-step is the one step ahead Newton Raphson maximization for α given the current values of the remaining quantities. Alternatively one may use more Newton Raphson iterations at the M-step but in practice this is not really helpful.

If some criterion is satisfied stop iterating else go back to E-step.

4.6. Neyman distribution

So far, all the mixing distributions considered were continuous distributions. The Neyman distribution is a discrete mixed Poisson distribution with many biological applications, see for example Douglas (1980). It is derived by considering a Poisson distribution with parameter kv where k follows itself a Poisson distribution with parameter μ. Then the resulting Neyman distribution has probability function given by

$$P(x|\mu,\nu) = \sum_{k=0}^{\infty} \frac{e^{-k\nu}(k\nu)^x}{x!} \frac{e^{-\mu}\mu^k}{k!} \tag{10}$$

for $x = 0, 1, \ldots$ and $\mu, \nu > 0$.

The following iterative scheme can be used for calculating the probabilities: $P(0|\mu,\nu) = \exp(-\mu + \lambda)$, where $\lambda = \mu \exp(-\nu)$ and

$$(x+1)\,P(x+1|\mu,\nu) = \nu\lambda \sum_{r=0}^{x} \frac{\nu^r}{r!} P(x-r|\mu,\nu)$$

for $x = 0, 1, \ldots$. This recursive scheme is the same as in Panjer (1981) taking into account the random sum representation of the Neyman distribution.

Douglas (1980) described the iterative scheme needed for ML estimation of the parameters. Sprott (1983) showed that one of the ML equations can be reduced to the first moment equation.

An EM type algorithm can be described as follows:

From the current estimates μ_{old} and ν_{old} the new estimates will be obtained as follows:

E-step: Calculate the pseudovalues

$$t_i = E(\theta_i|X_i) = \frac{(x_i+1)}{\nu_{\text{old}}} \frac{P(x_i+1|\mu_{\text{old}},\nu_{\text{old}})}{P(x_i|\mu_{\text{old}},\nu_{\text{old}})}$$

M-step: Find the new estimates as $\mu_{\text{new}} = \bar{t} = \sum_{i=1}^{n} t_i/n$ and $\nu_{\text{new}} = \bar{x}/\mu_{\text{new}}$.
If some criterion is satisfied stop iterating else go back to E-step.
There is no need for a numerical method and the scheme can be easily applied.

4.7. Poisson-Lognormal distribution

The Poisson lognormal distribution arises if we assume a lognormal mixing density. The probability function is given by

$$P\left(x|\mu,\sigma^2\right) = \frac{1}{\sigma\sqrt{(2\pi)\,x!}} \int_0^{\infty} (-\theta)\,\theta^{x-1} \exp\left(-\frac{(\log\theta-\mu)^2}{2\sigma^2}\right) d\theta$$

$x = 0, 1, \ldots$, $\mu \in \mathcal{R}$, $\sigma > 0$. Unfortunately the probability function cannot be simplified and hence the evaluation of the integral is necessary for calculating the probabilities. A thorough treatment of the distribution can be found in Shaban (1988). There are applications of the distribution in bibliometry (see, e.g. Stewart, 1994), species abundance data (e.g. Bulmer, 1974 among others). The distribution also arises in mixed effects Poisson regression when normal random effects are considered.

Estimation of the parameters via moment method is easy. For ML estimation one needs the probability function which is available only via numerical methods. ML estimation based on numerically approximated probabilities can be unstable. On the other hand, the ML estimates for the parameters of the lognormal distribution are rather easy to derive. One needs the quantities log X_i and $\log^2 X_i$. One can see that our general EM scheme applies. Namely the algorithm is as follows:

From the current estimates μ_{old} and σ_{old} the new estimates will be obtained as follows:

E-step: Calculate the pseudovalues

$$t_i = E(\log\theta_i | X_i) = \frac{\int_0^\infty \exp(-\theta)\,\theta^{x_i-1}\log\theta \exp\left(-\frac{(\log\theta - \mu_{\text{old}})^2}{2\sigma^2_{\text{old}}}\right)d\theta}{\int_0^\infty \exp(-\theta)\,\theta^{x_i-1}\exp\left(-\frac{(\log\theta - \mu_{\text{old}})^2}{2\sigma^2_{\text{old}}}\right)d\theta}$$

$$s_i = E\left(\log^2\theta_i \middle| X_i\right) = \frac{\int_0^\infty \exp(-\theta)\,\theta^{x_i-1}(\log\theta)^2 \exp\left(-\frac{(\log\theta - \mu_{\text{old}})^2}{2\sigma^2_{\text{old}}}\right)d\theta}{\int_0^\infty \exp(-\theta)\,\theta^{x_i-1}\exp\left(-\frac{(\log\theta - \mu_{\text{old}})^2}{2\sigma^2_{\text{old}}}\right)d\theta}$$

M-step: Find the new estimates as $\mu_{\text{new}} = \bar{t} = \sum_{i=1}^n t_i/n$ and $\sigma^2_{\text{new}} = \sum_{i=1}^n s_i/n - (\mu_{\text{new}})^2$.

If some criterion is satisfied stop iterating else go back to E-step.

Clearly the E-step has not closed form expressions and thus numerical approximations are needed. This can be either Monte Carlo approximation or numerical integrations. We will discuss in detail the case of Monte Carlo method in the sequel.

5. Application

In Table 1 the observed frequencies refer to the number of crimes for every month from 1982 until January 1993 (145 observations) in Greece. The data show overdispersion ($\bar{x} = 2.2413$, $s^2 = 3.3833$) making the assumption of a mixed Poisson distribution plausible. All the distributions described in section 4 were fitted to the data. As initial values for the iterations, the moment estimates for each distribution were used. We stopped the iterations if the relative change in the loglikelihood between two iterations was smaller than 10^{-10}. In Table 1 one can see the expected frequencies for each distribution. The estimated parameters and the maximized loglikelihood can be seen in Table 2.

Based on Table 1 one can deduce that the simple Poisson distribution gives very poor fit to our data set. The same is true for the Poisson-Lindley distribution which has a mode at 0, and a long right tail. On the other hand the remaining distribution give very good fit.

From Table 2 one can see that the fit of the Poisson inverse Gaussian, the negative binomial and the Poisson lognormal distributions is quite similar.

TABLE 1

Observed and expected frequencies for all the mixed Poisson distributions considered

x	observed	Poisson	Hermite	P-IG	P-Lindley	Neg. Bin	Neyman	P-Lognormal
0	21	15.42	23.18	22.93	39.12	23.55	29.38	22.85
1	41	34.55	32.90	35.92	31.51	35.19	31.99	35.91
2	32	38.72	32.95	32.73	23.52	32.15	28.96	32.83
3	16	28.93	24.67	23.07	16.77	23.16	21.67	23.17
4	19	16.21	15.58	14.07	11.59	14.44	14.36	14.09
5	8	7.27	8.51	7.87	7.83	8.17	8.69	7.83
6	4	2.71	4.16	4.17	5.20	4.30	4.89	4.12
7	1	0.87	1.85	2.14	3.41	2.15	2.59	2.10
8	2	0.24	0.76	1.07	2.21	1.03	1.30	1.05
9	1	0.06	0.29	0.53	1.42	0.48	1.15	.105
	χ^2	18.15	6.48	4.82	20.68	4.95	8.69	4.84
	df	6	5	5	6	5	5	5

TABLE 2

Estimated parameters for all the mixed Poisson distributions of table 1

distribution	parameters	loglikelihood
Poisson	$\hat{\lambda} = 2.131$	–281.485
Hermite	$\hat{\mu} = 2.2477$ $\hat{\sigma}^2 = 0.82846$	–275.4295
Geometric	$\hat{\lambda} = 0.4461$	–290.415
P-IG	$\hat{\delta} = 3.0959$ $\hat{\gamma} = 1.3812$	–274.4575
P.Lindley	$\hat{p} = 0.70122$	–284.2482
Neg.Binomial	$\hat{\alpha} = 4.49801$ $\hat{\beta} = 2.00183$	–274.5055
Neyman	$\hat{\mu} = 3.1041$ $\hat{v} = 0.72207$	–276.3262
P-Lognormal	$\hat{\mu} = 0.70$ $\hat{\sigma} = 0.20$	–274.4993

The geometric distribution is a special case of the negative binomial with $\alpha = 1$. The estimated value of α is far from 1 suggesting that the geometric distribution is not plausible. The standard error of $\hat{\alpha}$ is 1.675. One can also see the improvement on the loglikelihood between the geometric and the negative binomial models. Standard errors for all the models can be easily obtained through a bootstrap approach. The EM can facilitate this. In practice since good initial values are known the EM algorithm converges very quickly providing easily the standard errors.

It is interesting to see that for the Hermite distribution the assumed normal mixing distribution gives probability for θ taking a negative value near 0.001. For applying the EM algorithm this does not cause any problem at all. Recall, that for data of this kind, i.e. counts concentrated to small integers, it was expected that several distributions could fit well the data, as pointed in Douglas (1994).

6. Monte Carlo EM algorithm

So far, we have described mixed Poisson distributions for which the probability function is available at least via efficient recursive formulas. The exception was the Poisson lognormal distribution which does not have a probability function in closed form. We examined cases of either continuous or discrete mixing distributions, cases involving closed form M-steps or Newton Raphson steps inside the M-step. Now we will describe how we can avoid problems connected to the E-step rather than the M-step. These extensions of the typical EM algorithm are suitable for mixed Poisson distributions without closed form expressions for their probability function as, for example, the Poisson lognormal distribution.

6.1. Numerical Approximations and Stochastic versions of the EM

In many circumstances the direct application of the EM algorithm is not simple because the expectations involved at the E-step do not have closed form expressions. Then these expectations can be approximated numerically or they can be found via simulation based methods. For the former case, if the probability mass function of the mixed Poisson distribution is available, one may use (4) and a Taylor series expansion of the required expectation. On the contrary, if the probability mass function of the mixed Poisson distribution is not available, then one may use approximations proposed by Ong (1995) based on Taylor expansion of a special function of a gamma variate. Useful recursive formulas were given in Willmot (1993). Another standard numerical approach is numerical integration, see, e.g., Goutis (1993) and Aitkin (1996) for such a treatment.

The latter case leads to variants of the EM algorithm proposed in the literature. The Stochastic EM (SEM) (Celeux and Diebolt, 1985) and the Monte Carlo EM (MCEM) (Wei and Tanner, 1990) are two variants of the EM algorithm in which the E-step is based on simulation. Such approaches can be very helpful for improving the general behavior of the EM algorithm. First, the random perturbations due to simulation can prevent the algorithm from stopping in local maxima or saddle points improving also the convergence rate. Moreover, the underlying dynamics help the algorithm to reach the target value after a comparatively small number of iterations. Finally, the statistical considerations stemming from the simulated step allows for obtaining an idea of the variability of the estimates. For both algorithms, after some iterations the estimates converge to a stationary point and thus ergodic means can be used as an estimate of the parameter. The convergence of the estimates can be checked via convergence diagnostics developed for Markov Chain Monte Carlo methods (see, Cowles and Carlin, 1996).

Suppose that at the E-step, the conditional expectation $E(h(\theta)\,|\,X, \varphi^{(k)})$ is needed. Using the first approach, the Monte Carlo EM algorithm, one generates m values $h(\theta^{(i)})$, $i, \ldots, m$ from the conditional density $f(\theta | x_i, \varphi^{(k)})$ and approximates the conditional expectation by

$$E\left(h\left(\theta_i\right) \middle| X_i, \varphi^{(k)}\right) \approx \frac{\sum_{i=1}^{m} h\left(\theta^{(i)}\right)}{m}$$

If $m = 1$ then the SEM algorithm is derived. If m is very large the MCEM algorithm works approximately like the EM, and thus it has all the pros and the cons of the standard EM. However, the choice of large m results in a very slow algorithm and in practice is useless. Both of the algorithms do not have the monotonic behavior of the EM algorithm due to the random perturbations of the E-step.

Note that the above schemes do not require knowledge of the probability mass function. It suffices to be able to simulate from the posterior density $g(\theta | x)$. The general algorithm for simulating from the posterior distribution can be used (see, e.g., Tanner, 1996). It is a rejection method. For the Poisson distribution this scheme for simulating from the posterior distribution $g(\theta | x_i, \varphi^{(k)})$ is as follows:

- Generate θ from $g(\theta | \varphi^{(k)})$
- Accept this value if $U < \frac{e^{-\theta}\theta^{xi}}{e^{-xi}x_i^{x\,i}}$ where U is a uniform variate

This algorithm has the undesirable feature that it is very slow for posterior distributions when the value of x is at the right tail of the distribution. However in some cases more efficient methods can be found for simulating from the posterior density. There is a wealth of such procedures used in the Bayesian setting for Markov Chain Monte Carlo approaches. Alternatively one may use Markov Chain Monte Carlo method (McCullogh, 1997) or importance sampling (Booth and Hobert, 1999) in order to simulate from the conditional distribution.

There are two remaining issues related to the MCEM algorithm. The extensive sampling at each step can make the algorithm very slow and clearly the MCEM algorithm does not have the good monotonic properties of the EM algorithm.

The first issue has to do with the efficient use of the samples taken at each step, namely the choice of good values of m so as to improve the speed of the algorithm. The general strategy is to use small m for the first iterations, since the algorithm can reach the area of the maximum relatively easily and then to increase m in order to be able to maximize the likelihood near the maximum. Booth and Hobert (1999) provide an interesting treatment of the problem, see also Levine and Casella (2001).

The second issue relates to the first one and it is much more difficult to check. Usually, the MCEM stabilizes after some iterations to a region where the maximum exists and thus some more iterations are sufficient in order to take good estimates. The algorithm converges under suitable regularity conditions (see, e.g. Chan and Ledolter, 1995).

6.2. Illustration

6.2.1. The negative binomial case

It is well known that in the case of a negative binomial distribution the conditional distribution of $\theta|x$ is a gamma density with parameters $\alpha + x$ and $\beta + 1$. Thus simulating form the conditional distribution is straightforward. Therefore, the E-step implies that the expectations are estimated through simulating m values $\theta^{(j)}$, $j = 1,\dots,m$, from a $Gamma(\alpha + x,\ \beta + 1)$ density and using

$$t_i = E(\theta_i|x_i) \approx m^{-1}\sum_{j=1}^{m}\theta^{(j)} \text{ and } s_i = E(\log\theta_i|x_i) \approx m^{-1}\sum_{j=1}^{m}\log\theta^{(j)}$$

Then the M-step given in section 4.5 applies.

Clearly the negative binomial case is used for illustration. Since we know the exact form of the algorithm this can serve as a basis to illustrate the dynamics of the algorithm.

6.2.2. The Poisson lognormal case

For the Poisson lognormal distribution the proposed algorithms are much more useful since no closed form expressions can be used. The algorithm is the following:

E-step: Simulate values $\theta^{(i)}$, $i = 1,\dots,m$ from the posterior density $g(\theta|x_i)$. Then, estimate the posterior expectations

$$t_i = E(\log\theta_i|x_i) = m^{-1}\sum_{j=1}^{m}\log\theta^{(j)} \text{ and } s_i = E\left(\log^2\theta_i\middle|x_i\right) = m^{-1}\sum_{j=1}^{m}\log^2\theta^{(j)}.$$

M-step: Update the parameters of the lognormal distributions as

$$\mu_{\text{new}} = n^{-1}\sum_{i=1}^{n}t_i \text{ and } \sigma^2_{\text{new}} = n^{-1}\sum_{i=1}^{n}s_i - \mu_{\text{new}}{}^2$$

It is very appealing that this scheme does not require at all knowledge of the probability function of the Poisson lognormal distribution.

6.2.3. Numerical Example

We used the MCEM described above to the data set considered in section 5. For the negative binomial the exact ML estimates are known, so we will use the results of the MCEM to show the usefulness of the algorithm.

In figure 1 one can see the history of the MCEM using different choices of m, namely $m = 1$, 10, 100 and finally $m = 15j$, where j is the iteration number. There are some interesting points arising from figure 1. First of all the variability

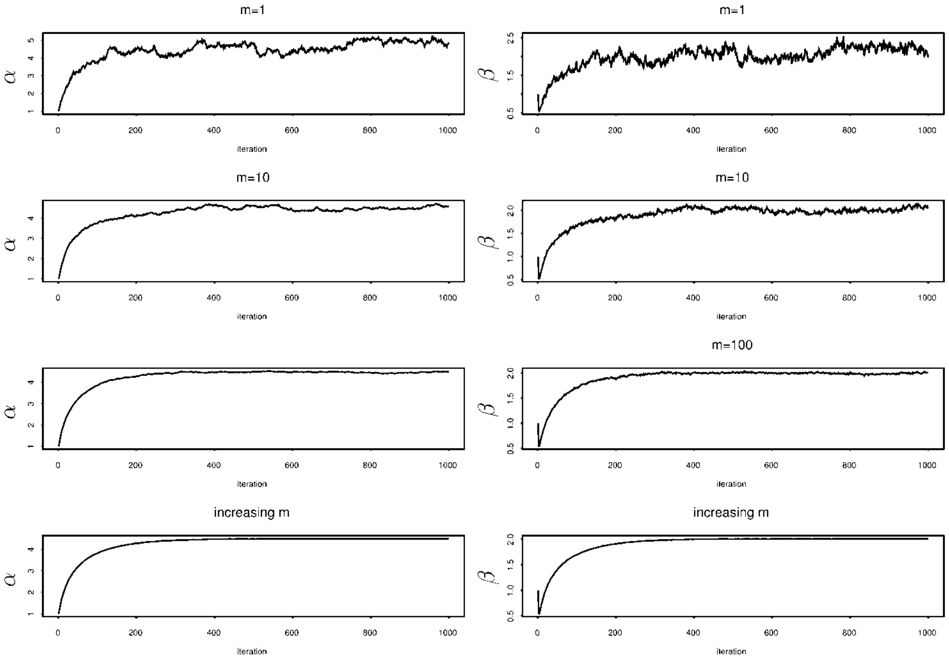

FIGURE 1: MCEM for the negative binomial model, using different choices of m.

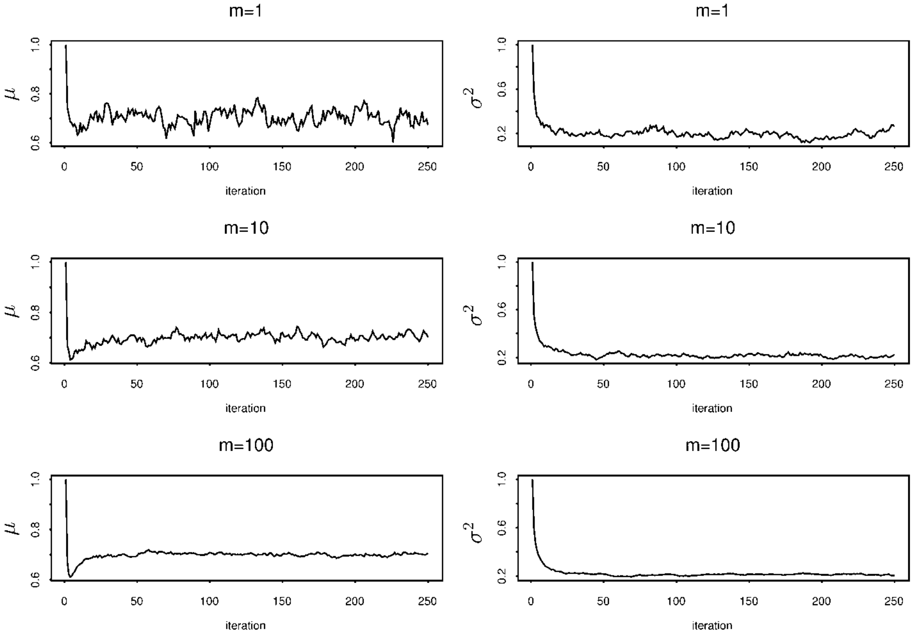

FIGURE 2: MCEM for the Poisson lognormal model, using different choices of m.

clearly depends on the choice of m. If one looks for the case when $m = 1$, the algorithm approaches the area of the estimate and then perturbates around this point. This perturbation is much smaller when m increases. Another interesting point is that whatever the choice of m the algorithm approaches the estimate at the same speed. The choice of m has an effect on how much the estimates perturbate around the ML estimate. The strategy with increasing m locates the maximum almost with the same number of iterations as the standard EM algorithm. The number of iterations can be also seen in the x-axis.

From the practical point of view one can run the algorithm with increasing m and to examine whether the chain has stabilized and, then, to obtain an ergodic mean of the chain as an estimate. Usually the chain has quite large autocorrelations and this must be taken into account if an estimate of the variance is to be calculated.

A similar plot for the case of the Poisson lognormal distribution can be seen in figure 2. The algorithm described in section 6.1 was used to simulate form the conditional posterior density of θ. The behavior of the algorithm is the same as the one for the negative binomial case.

7. Other Discrete mixture distributions

In this section we briefly extend the EM approach to other mixtures of discrete distributions, like mixtures of the geometric distribution and the binomial distribution.

7.1. The Yule distribution

Consider the geometric distribution defined in (5), and let the parameter be $\lambda = \theta(1-\theta)^{-1}$. Consider also the beta Type I distribution, denoted as $BetaI(\alpha,\beta)$ with density function

$$g(\theta) = \theta^{\alpha-1}(1-\theta)^{\beta-1}/B(\alpha,\beta), \quad \alpha,\beta > 0,\ 0 \le \theta \le 1, \tag{11}$$

where $B(\alpha,\beta)$ is the usual beta function. If X conditional on θ follows a geometric distribution with parameter $\theta(1-\theta)^{-1}$ and θ follows a beta Type I $(\alpha,1)$ distribution with density given by $g(\theta) = \alpha\theta^{\alpha-1}$, then unconditionally X follows a Yule distribution (see Johnson et al, 1992). The probability function of the Yule distribution is given by

$$P(x) = \frac{\alpha\Gamma(\alpha+1)\,x!}{\Gamma(a+x+2)}, \tag{12}$$

$x = 0,1\ldots,\ \alpha > 0$. For a sample of size n, θ_i, $i = 1,\ldots,n$ from the $g(\theta)$ density the MLE of the parameter α is easily calculated as $\alpha = -n^{-1}\sum_{i=1}^{n}\log\theta_i$. Note that if $g(\theta)$ is a beta Type I $(\alpha,1)$ distribution then the density of $\theta|x_i$ is a beta Type I distribution with parameters $(\alpha+1,\ x_i+1)$.

Hence the EM algorithm can be described as:

E-step: Calculate $t_i = E(\log\theta_i | X_i) = \Psi(\alpha_{\text{old}} + 1) - \Psi(\alpha_{\text{old}} + 2 + X_i)$

M-step: Update the parameter using $\alpha_{\text{new}} = -n^{-1}\sum_{i=1}^{n} t_i$

The Yule distribution was applied to the data considered in Table 1. The fit was not good as expected since the Yule distribution is J-shaped with a large tail, something not apparent in the data set.

7.2. The beta-binomial distribution

Binomial mixtures arise naturally if one assumes that the probability p of success at each trial is not constant but it varies according to a mixing distribution $g(p)$. A common choice is the beta distribution as mixing distribution giving rise to the beta-binomial distribution.

The beta-binomial (BB) distribution has probability function given by

$$P(x|\alpha,\beta) = \binom{N}{x}\frac{B(\alpha + x, N + \beta - x)}{B(\alpha,\beta)} \tag{13}$$

where N is the number of trials. The mean and the variance of the BB distribution are $N\pi$ and $N\pi(1-\pi)(N\varphi + 1)(1 + \varphi)^{-1}$ where $\pi = \alpha(\alpha + \beta)^{-1}$ and $\varphi = (\alpha + \beta)^{-1}$.

ML estimation for the parameters of the BB distribution is not an easy task, since the derivatives of the beta function are involved. Griffiths (1973) described ML estimation using numerical techniques. Tripathi *et al.* (1994) described other methods of estimation.

When trying to find ML estimates from a random sample of beta random variables X_i one needs the quantities $\log X_i$ and $\log(1 - X_i)$. Thus at the E-step one has to find the quantities $E(\log p_i | X_i = x_i)$ and $E(\log(1 - p_i) | X_i = x_i)$ using the current values of the parameters. It is well known from Bayesian statistics that if the density $f(p_i)$ is a beta density with parameters α and β then the density $f(p_i | X_i = x_i)$ is again a beta density with parameters $\alpha + x_i$ and $N + \beta - x_i$ respectively. Thus the quantities needed can be easily calculated using the formulas for the $E(\log X)$ and $E(\log(1 - X))$ for a beta variate X. The derivation of these expectations can be found in the Appendix.

Thus the EM algorithm can be described as:

From the current estimates α_{old} and β_{old} the new estimates will be obtained as follows:

E-step: Calculate

$$t_i = E(\log p_i \,|\, X_i = x_i) = \Psi(\alpha_{\text{old}} + x_i) - \Psi(\alpha_{\text{old}} + \beta_{\text{old}} + N)$$

$$s_i = E(\log(1 - p_i \,|\, X_i = x_i) = \Psi(\beta_{\text{old}} + N - x_i) - \Psi(\alpha_{\text{old}} + \beta_{\text{old}} + N)$$

for $i = 1,\ldots,n$.

M-step: Make an one-step ahead Newton Raphson iteration for ML estimation of a beta density using the expectations of the E-step. To do so calculate

$$\bar{t} = \frac{\sum_{i=1}^{n} t_i}{n} \quad \text{and} \quad \bar{s} = \frac{\sum_{i=1}^{n} s_i}{n}$$

and then update the estimates as

$$\alpha_{\text{new}} = \alpha_{\text{old}} - \frac{\Psi(\alpha_{\text{old}}) - \Psi(\alpha_{\text{old}} + \beta_{\text{old}}) - \bar{t}}{\Psi_3(\alpha_{\text{old}}) - \Psi_3(\alpha_{\text{old}} + \beta_{\text{old}})}$$

and

$$\beta_{\text{new}} = \beta_{\text{old}} - \frac{\Psi(\beta_{\text{old}}) - \Psi(\alpha_{\text{new}} + \beta_{\text{old}}) - \bar{s}}{\Psi_3(\beta_{\text{old}}) - \Psi_3(\alpha_{\text{new}} + \beta_{\text{old}})}.$$

Note that alternative schemes can be constructed by trying to directly maximize the beta likelihood at the M-step. It was found that such a maximization could delay the algorithm since at each M-step several iterations are needed to maximize the complete likelihood.

The algorithm is easy to be programmed in many statistical packages as it needs only the specification of the digamma and trigamma functions. Neither matrix inversions nor other numerical techniques are needed. A Monte Carlo EM version for the beta binomial is described in Booth and Caffo (2002).

As an application consider the data in Table 3. The data concern the number of passed courses for a class of 65 students from the first year of the Department of Statistics of Athens University of Economics. This class attended 8 courses during the year. The number of successful examinations were recorded. The binomial distribution with $\hat{p} = 0.65$ had a very poor fit. This was expected since it is not reasonable to consider the probability of success p to be constant for all the courses. It seems natural to consider that the courses have different difficulty and thus the probability of success varies. Assuming that this probability varies according to a beta distribution the BB distributions arises. Fitting the BB distribution the estimates obtained were $\hat{\alpha} = 1.825$ and $\hat{\beta} = 0.968$. The fit can be seen in Table 3. The χ^2 statistic had a value 1.45 with 6 degrees of freedom (no grouping was made) which shows a very good fit. Note also that the loglikelihood for the beta-binomial model was – 134.75 instead of –168.82 for the simple binomial model indicating large improvement. Thus it is reasonable to conclude the varying difficulty of the courses.

TABLE 3

DATA CONCERNING THE NUMBER OF PASSED COURSES FOR A CLASS OF 65 STUDENTS AT THE DEPT OF STATISTICS, ATHENS UNIVERSITY OF ECONOMICS ($N = 8$).

x	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
observed	1	4	4	8	9	6	8	12	13
expected	1.80	3.28	4.65	5.97	7.25	8.51	9.78	11.11	12.65

8. Concluding Remarks and Discussion

In this paper, ML estimation of the parameters of mixed Poisson distributions were treated in detail, using an EM type algorithm that uses certain properties of this family of distributions. Some other mixtures of discrete distributions were discussed, too. It is clear that the approach developed can be expanded in several ways. First of all, every distribution arising as a mixture of some other distribution can be treated in this way. Of course in certain circumstances this approach can be less efficient than other methods but the EM algorithm itself provides insight into the estimation task. For example, in the mixed Poisson case, quantities like $E(\theta|x_i)$ are important for further inference, like experienced rate in actuarial practice. These quantities are byproducts of the algorithm. Moreover, this quantity characterizes the risk of the i-th individual and then it can be used for various goals as for example Empirical Bayes methods. The approach can be also expanded to cover bivariate and multivariate distributions. For example Munkin and Trivedi (1999) used simulated ML approach for a mixture of a bivariate Poisson distribution with a lognormal mixing density. The proposed EM algorithm approach is clearly easier than their simulated ML approach.

Secondly, random effects models can be estimated in a similar manner. In Karlis (2001) negative binomial regression and Poisson-inverse Gaussian regression was treated. The present paper aims at providing more details on the algorithm presented there.

References

Aitkin, M. (1996) A General Maximum Likelihood Analysis of Overdispersion in Generalized Linear Models. *Statistics and Computing*, **6**, 251-262.

Booth, J.G. and Caffo, B.S. (2002) Unequal sampling for Monte Carlo EM algorithms. *Computational Statistics and Data Analaysis*, **39**, 261-270.

Booth, J.G. and Hobert, J.P. (1999) Maximizing generalized linear mixed model likelihood with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B*, **61**, 265-285.

Bohning, D. (1999) *Computer assisted analysis of mixtures and applications in meta-analysis, disease mapping and others*. CRC Press, New York.

Bulmer M. (1974) On fitting the Poisson-lognormal distribution to species-abundance data. *Biometrics*, **30**, 101-110.

Carlin, B.P., and Louis, T.A. (1996) *Bayes and Empirical Bayes Methods for Data Analysis*. Great Britain: Chapman and Hall.

Celeux, G. and Diebolt, J. (1985) The SEM Algorithm, a Probabilistic Teacher Algorithm Derived from the EM Algorithm for the Mixture Problem. *Computational Statistics Quarterly*, **2**, 73-82.

Chan, K.S. and Ledolter, J. (1995) Monte Carlo EM estimation for time series models involving counts. *Journal of the American Statistical Association*, **90**, 242-252.

Cowles, M.K. and Carlin, B.P. (1996) Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, **91**, 883-904.

Dempster, A.P., Laird, N.M. and Rubin, D. (1977) Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society*, **B 39**, 1-38.

Douglas, J.B. (1980) *Analysis With Standard Contagious Distributions*. Statistical distributions in scientific work series, Vol 4, ICPH, Maryland, USA.

Douglas, J.B. (1994) Empirical Fitting of Discrete Distributions. *Biometrics*, **50**, 576-579.

Goutis, C. (1993) Recovering Extra Binomial Variation. *Journal of Statistical Computation and Simulation*, **45**, 233-242.

Grandell, J. (1997) *Mixed Poisson Processes.* Chapman and Hall/CRC Statistics and Mathematics, New York.

Griffiths, D.A. (1973) Maximum Likelihood Estimation for the Beta Binomial Distributions and an Application to the Household Distribution of the Total Number of Cases of a Disease. *Biometrics*, **29**, 637-648.

Greenwood, M. and Yule, G. (1920) An Inquiry Into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurence of Multiple Attacks of Disease Or of Repeated Accidents. *Journal of the Royal Statistical Society*, **A 83**, 255-279.

Johnson, N., Kotz, S. and Kemp, A.W. (1992) *Univariate discrete distributions*, Willey-New York.

Jorgensen, B. (1982) The generalized Inverse Gaussian distribution. *Lecture Notes in Statistics*, **9**, Springer-Verlag.

Karlis, D. (1998) Estimation and hypothesis testing problems in Poisson mixtures. *Phd Thesis, Department of Statistics, Athens University of Economics*.

Karlis, D. (2001) Exact ML estimation for Mixed Poisson Regression Models. *Statistical Modelling: An International Journal*, **1**, 305-319.

Kemp, C.D. and Kemp, A.W. (1965) Some Properties of the Hermite Distribution. *Biometrika*, **52**, 381-394.

Kemp, A.W. and Kemp, C.D. (1966) An Alternative Derivation of the Hermite Distribution, *Biometrika*, **53**, 627-628.

Levine, R.A. and Casella, G. (2001) Implementations of the Monte Carlo EM Algorithm. *Journal of Computational and Graphical Statistics*, **10**, 422-439.

Lindsay, B. (1995) *Mixture Models: Theory, Geometry and Applications*. Regional Conference Series in Probability and Statistics, Vol 5, Institute of Mathematical Statistics and American Statistical Association.

McCulloch, P. (1997) Maximum Likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association*, **92**, 162-170.

McLachlan, G. and Peel, D. (2000) *Finite mixture models*, Wiley, New York.

Meng, X.L. and Rubin, D. (1993) Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika*, **80**, 267-278.

Munkin, M.K. and Trivedi, P.K. (1999) Simulated Maximum Likelihood Estimation of Multivariate Mixed-Poisson Regression Models, with Application. *Econometrics Journal*, **2**, 29-48.

Ong, S.H. (1995) Computation of Probabilities of a Generalised log-series and Related Distributions. *Communication in Statistics – Theory and Methods*, **24**, 253-271.

Panjer, H. (1981) Recursive Evaluation of a Family of Compound Distributions. *ASTIN Bulletin*, **18**, 57-68.

Piegorsch, W.W. (1990) Maximum Likelihood Estimation for the Negative Binomial Dispersion Paramater. *Biometrics*, **46**, 863-867.

Sankaran, M. (1970) The Discrete Poisson-Lindley Distribution. *Biometrics*, **26**, 145-149.

Sapatinas, T. (1995) Identifiability of mixtures of power-series distributions and related characterizations, *Annals of the Institute of Statistical Mathematics*, **47**, 447-459.

Sichel, H.S. (1974) On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society*, **A 137**, 25-34.

Sichel, H.S. (1982) Asymptotic efficiencies of three methods of estimation for the inverse gaussian-Poisson distribution. *Biometrika*, **69**, 467-472.

Shaban, S.A. (1988) Poisson-Lognormal Distributions. In E.L. Crow and K. Shimizu (eds), *Lognormal Distributions: Theory and Applications*, 195-210, New York, Marcel and Dekker.

Sprott D. (1983) Estimating the parameters of a convolution by the Maximum Likelihood. *Journal of the American Statistical Association*, **78**, 457-460.

Stewart, J.A. (1994) The Poisson-Lognormal Model for Bibliometric/Scientometric Distributions. *Information Processing and Management*, **30**, 239-251.

Tanner, M.A. (1996) *Tools for statistical inference: Methods for the Exploration of Posterior distributions and Likelihood Functions*, 3rd edition, Springer, New York.

TITTERINGTON D.M, SMITH A.F.M. and MAKOV U.E. (1985) Statistical analysis of finite mixtures distributions. Willey and sons, New York.

TRIPATHI, R., GUPTA, R. and GURLAND, J. (1994) Estimation of the Parameters in the Beta Binomial Model. *Annals of the Institute of Statistical Mathematics*, **46**, 317-331.

WANG, Y. (1996) Estimation Problems for the Two-Parameter Negative Binomial Distribution. *Statistics and Probability Letters*, **26**, 113-114.

WEI, G.C.G. and TANNER, M.A. (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, **85**, 699-704.

WILLMOT, G. (1993) On Recursive Evaluation of Mixed Poisson Probabilities and Related Quantities. *Scandinavian Actuarial Journal*, 101-113.

APPENDIX

It holds that

$$\frac{\Gamma(\alpha)}{\lambda^{\alpha}} = \int_0^{\infty} x^{\alpha-1} \exp(-\lambda x)\, dx$$

and hence by differentiating both sides with respect to α yields

$$\frac{\Gamma'(\alpha)\lambda^{\alpha} - \lambda^{\alpha}\log(\lambda)\Gamma(\alpha)}{\left(\lambda^{\alpha}\right)^2} = \frac{\Gamma(\alpha)[\Psi(\alpha) - \log(\lambda)]}{\lambda^{\alpha}} = \int_0^{\infty} \log x \; x^{\alpha-1} \exp(-\lambda x)\, dx$$

where $\Psi(\alpha)$ denotes the digamma function defined as

$$\Psi(\alpha) = \frac{\partial \log \Gamma(\alpha)}{\partial \alpha} = \frac{1}{\Gamma(\alpha)} \frac{\partial \Gamma(\alpha)}{\partial \alpha}$$

Thus for a Gamma variate it holds that

$$E(\log(X)) = \Psi(\alpha) - \log(\lambda)$$

With a similar argument it can be seen that for the Beta distribution given in (11) it holds that

$$E(\log X) = \Psi(\alpha) - \Psi(\alpha+\beta) \;\text{ and }\; E(\log(1-X)) = \Psi(\beta) - \Psi(\alpha+\beta)$$

DIMITRIS KARLIS
Department of Statistics
Athens University of Economics and Business
76, Patission str, 10434, Athens, Greece
Email: karlis@aueb.gr