© Board of the Foundation of the Scandinavian Journal of Statistics 2005. Published by Blackwell Publishing Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA Vol 32: 115–131, 2005

Asymptotic Normality in Mixtures of Power Series Distributions

DANKMAR BÖHNING

Institute for Social Medicine, Epidemiology and Health Economics, Charité Medical School Berlin

VALENTIN PATILEA

CREST-ENSAI

ABSTRACT. The problem of estimating the individual probabilities of a discrete distribution is considered. The true distribution of the independent observations is a mixture of a family of power series distributions. First, we ensure identifiability of the mixing distribution assuming mild conditions. Next, the mixing distribution is estimated by non-parametric maximum likelihood and an estimator for individual probabilities is obtained from the corresponding marginal mixture density. We establish asymptotic normality for the estimator of individual probabilities by showing that, under certain conditions, the difference between this estimator and the empirical proportions is asymptotically negligible. Our framework includes Poisson, negative binomial and logarithmic series as well as binomial mixture models. Simulations highlight the benefit in achieving normality when using the proposed marginal mixture density approach instead of the empirical one, especially for small sample sizes and/or when interest is in the tail areas. A real data example is given to illustrate the use of the methodology.

Key words: asymptotic normality, identifiability, mixture models, non-parametric maximum likelihood, power series distributions

1. Introduction

Mixtures of distributions are commonly used in a wide range of applications, see Titterington *et al.* (1985), Lindsay (1995) and McLachlan & Peel (2000) for comprehensive descriptions of the mixture landscape. In particular, mixtures of discrete distributions represent a popular tool for analysing count data. Herein, the interest is focused on mixtures of power series distributions, also called linear exponential distributions. To estimate such mixtures we consider the non-parametric maximum likelihood approach, which is known to be appropriate for applications where only a little information about the true mixing distribution is available.

Consider X a discrete random variable distributed according to π_{Q_0}, a mixture of a given family $\{\pi_\theta, \theta \in \Theta\}$ of discrete distributions with unknown mixing distribution Q_0, that is $\pi_{Q_0} = \int_\Theta \pi_\theta Q_0(\mathrm{d}\theta)$. Suppose that we observe an i.i.d. sample distributed according to π_{Q_0}. We are interested in estimating probabilities like

$$\pi_{Q_0,J} = \sum_{k \in J} \pi_{Q_0,k} = P(X \in J), \tag{1}$$

where J is any finite subset of the support of the observed variable and $\pi_{Q_0,k} = P(X = k)$.

Let $\hat{Q}$ and $\hat{\pi}$ denote the non-parametric maximum likelihood estimator of Q_0 and the corresponding mixture, respectively. We investigate the asymptotic distribution of $\hat{\pi}_J$, the estimator of $\pi_{Q_0,J}$ yielded by the estimated mixture $\hat{\pi}$. This problem has been analysed by Lambert & Tierney (1984) in the case of Poisson mixtures. Their main result shows that the difference between $\hat{\pi}_J$ and the proportion of observations belonging to J is asymptotically negligible, that is of order $o_{\mathbf{P}}(n^{-1/2})$ where n is the sample size. If Z_n, $n \geq 1$, is a sequence of random variables, we say that Z_n is of order $o_{\mathbf{P}}(k_n)$ if Z_n/k_n converges to zero, in probability. Next, the asymptotic normality of $\hat{\pi}_J$ follows from a classical central limit theorem. Lambert & Tierney assumed that the support of the true mixing distribution is contained in [0, M], a *known* compact interval, and that the non-parametric maximum likelihood estimation is restricted to mixing distributions with the support contained in this interval. Below, we show that their proof does not apply to power series distributions with a finite radius of convergence, except for the case where M is the upper limit of the support of Q_0.

Here, we extend the results of Lambert & Tierney (1984) on the asymptotic equivalence between $\hat{\pi}$ and the empirical proportions to more general power series distributions (PSDs) compound families. Our results apply also to families with finite radius of convergence such as negative binomial, geometric and logarithmic series distributions. The case of compound families with finite support (e.g. binomial laws) is also studied. The estimator of $\pi_{Q_0,J}$ that we study in this paper is the one given by the unrestricted non-parametric maximum likelihood, that is the maximum is taken over all mixing probability measures.

To obtain our results for the case of compound families with infinite support we assume that Q_0 is concentrated on some *unknown* compact interval [0, M]. Moreover, the support of Q_0 has to be an infinite set. It is also shown that, in general, the difference between $\hat{\pi}$ and the empirical proportions is no longer negligible if Q_0 has a finite support, that is when the law of the observations is a finite mixture. Whether the quantities $\hat{\pi}_J$ with J finite still have normal asymptotic behaviour is not known. Simulation results indicate that, even if Q_0 is discrete, $\hat{\pi}_J$ is still asymptotically normal, at least for some sets J.

In the case of PSD compound families with finite support, almost surely, $\hat{\pi}$ is equal to the empirical proportions provided that the true mixture is an interior point of the set of all mixtures of the model and the sample size is sufficiently large. If π_{Q_0} is on the boundary of the set of all mixtures, the difference between $\hat{\pi}$ and the empirical proportions is no longer negligible.

The assumptions on Q_0 we impose for the asymptotic normality results ensure that the empirical proportions represent an efficient estimator for the theoretical probabilities, see e.g. van der Vaart (1998, section 25.5.2). From our results it follows that the probability mass estimators obtained from $\hat{\pi}$ are also efficient. Even if the estimated mixture is asymptotically equivalent to the naive empirical distribution, this latter probability distribution is not a mixture and thus it is useless for the interpretation of the latent structure underlying the model. Moreover, whereas the simple estimator of $P(X \in J)$ based upon the empirical proportions is frequently difficult to use (not to say useless) when the sample sizes are small and/or interest is in the tail areas, the estimator based upon the marginal mixture density $\hat{\pi}$ is of considerable practical value.

The paper is organized as follows. We end this section with a list of possible applications for the mixtures we consider herein and where the probabilities $P(X \in J)$, J finite or some of their transformations represent quantities of interest. In section 2, some definitions and basic facts on the families of power series distributions are recalled. Moreover, a mixture identification result is proved under mild conditions. Section 3 recalls some results on the non-parametric maximum likelihood estimator and contains the orders of some χ^2-type norms as introduced by Lambert & Tierney (1984). The asymptotic normality results are proved in section 4. Section 5 contains the output of a simulation experiment using mixtures of Poisson and geometric distributions. The behaviour of non-parametric maximum likelihood and empirical estimators are analysed and compared. There is considerable evidence that the normal approximation is improved when using the proposed marginal mixture density approach in comparison with the conventional empirical one. Finally, a real data example is considered that demonstrates the practical value of the proposed method.

1.1. Examples of application areas

Non-life insurance. A rating model in insurance computes risk premiums, which are estimations of risk levels. These levels depend crucially on the number of claims. A usual assumption is that each individual has its own Poisson distribution for the number of claims. The heterogeneity of the population of policy holders makes the distribution of the observed frequency of claims to be a Poisson mixture (Simar, 1976). The probabilities of observing k claims, $k = 0, 1,...$, represent quantities of interest for risk management.

Biology and medicine. Most countries are collecting records that allow the provision of nationwide vital statistics, counting births and deaths, and categorizing deaths according to cause of death. Hasselblad (1969) mentions the famous Times-Count-Data set in which the daily death notices of the London newspaper *The Times* were collected over the 3-year period 1910–12 (see also Titterington *et al.*, 1985). Böhning *et al.* (1992) study the incidence data in a population of pre-school children in north-east Thailand for which the frequency of acute respiratory infection is counted over a 3-year period. For all, these kinds of count data mixtures of Poisson distributions are often used for modelling the count distribution. Specific count data arise in fertility studies providing interest in the geometric distribution and mixtures thereof (e.g. Böhning, 2000).

Capture–recapture studies. In ecology, the size of an animal population is frequently modelled using a truncated count distribution where the counts represent the capture–recapture history of each of the animals. Here, the probability for a zero-count (e.g. the probability of an animal never being captured in any of the capture samples) is of intrinsic interest, as this probability allows the adjustment of the observed sample size by the inverse of the probability of not observing a zero-count to achieve the Horvitz–Thompson estimator of the size of the population (e.g. Kendall & Stuart, 1991). Clearly, the simple proportions are not able to provide a solution to this *prediction problem*. Simple count distributions such as the Poisson or the binomial provide a starting point for modelling. Mixtures of truncated count distributions such as truncated Poisson or binomial distributions represent a more flexible tool for coping with population heterogeneity (e.g. Mao & Lindsay, 2002b).

Reliability. In most settings involving failure data, the population under study is not homogenous. Mixture models provide a natural answer to this problem, in particular for discrete failure observations. Discrete failure data arise in various common situations in reliability where chronological time is not the best scale on which to describe lifetime, see Gupta *et al.* (1997) and Shaked *et al.* (1994). For example, when a piece of equipment operates on demand, the number of operations successfully completed might be more important than the age at failure. Quantities of interest in such a framework are the survivor function $P(X > k)$ and the hazard rate function

$$\lambda(k) = \frac{P(X = k)}{P(X \geq k)}$$

computed at a point k in the support of a lifetime X.

Another situation where discrete data appear in reliability is when a device can be monitored only once per time period. In such a case the observation consists of the number of time periods completed prior to failure. Consider, for instance, that the true distribution of the lifetime Y is a mixture of exponentials, that is the density of Y is

$$f_{\tilde{Q}}(t) = \int_{(0,\infty)} \lambda \mathrm{e}^{-\lambda t} \tilde{Q}(\mathrm{d}\lambda), \qquad t \geq 0,$$

© Board of the Foundation of the Scandinavian Journal of Statistics 2005.

with $\tilde{Q}(\lambda)$ the mixing distribution. In this case, by Fubini's theorem we obtain

$$P(Y \in [k, k+1)) = \int_{[k,k+1)} f_{\tilde{\mathrm{Q}}}(t)\mathrm{d}t = \int_{(0,1)} (1-\theta)\theta^k Q_0(\mathrm{d}\theta), \qquad k = 0, 1, \ldots,$$

where Q_0 is the distribution obtained from $\tilde{Q}$ after the change of variable $\theta = \mathrm{e}^{-\lambda}$. In other words, observing Y at times $k = 0,1,\ldots$ is like observing a variable with distribution as a mixture of negative binomials.

Empirical Bayes. The Bayesian approach to inference requires a prior distribution for the model parameters. If this distribution, say Q, is given, the best mean-squared error estimator is the posterior mean. In the case of a Poisson model we have

$$E[\theta \mid X = k] = \frac{(k+1)\pi_{Q,k+1}}{\pi_{Q,k}}, \qquad k = 0, 1, \ldots, \tag{2}$$

that is, the Bayes rule can be written in terms of the marginal distribution π_Q and the value taken by X. Note that this property is shared by all the discrete power series distributions mixture models. If Q is not known, one may estimate $\pi_{Q,k}$s from the observed data. This is the empirical Bayes (EB) approach (e.g. Carlin & Louis, 1996). In the parametric version of the EB methodology, the prior Q is assumed to belong to a specified family indexed by unknown parameters. These parameters are estimated from the observed data through the marginal distribution, for example, by maximum likelihood. See, for instance, Martz *et al.* (1996) for some applications of the parametric EB method for binomial, geometric and Poisson sampling models commonly used for studying the reliability of the nuclear plant equipment. Robbins (1955) proposed the purely non-parametric approach where the marginal probabilities $\pi_{Q,k}$ are estimated by the corresponding empirical frequencies. An appealing compromise between parametric and purely non-parametric approaches for EB is obtained when Q is estimated by non-parametric maximum likelihood from data (see Carlin & Louis, 1996, section 3.2.3, for a comparison of various EB estimators). For instance, Mao & Lindsay (2002a) use the EB estimator (2) with Q estimated by non-parametric maximum likelihood in a Poisson sampling model in order to estimate a conditional prediction function for the number of expressed genes that have not been observed in the initial sample but can be identified in an additional sample.

2. Identifiability of mixtures of power series distributions

Consider a power series $a(\theta) = \sum_{k \geq 0} a_k \theta^k$, $a_k \geq 0$, $k = 0,1,\ldots$, and let R be its radius of convergence. Denote by Θ its domain of convergence on the non-negative half-line, that is $\Theta = [0, R]$ if $a(R)$ is finite and $\Theta = [0, R)$ otherwise. Let $\mathbb{K} = \{k{:}a_k > 0\}$ and

$$\pi_{\theta,k} = a_k \theta^k a(\theta)^{-1}, \; k \in \mathbb{K}, \; \theta \in \Theta,$$

be the compound PSD with support $\mathbb{K}$. Such distributions are also called discrete linear exponential distributions. Two types of compound families can be distinguished depending on whether the support $\mathbb{K}$ is finite or not. Some common examples of PSDs with infinite support are Poisson [$a(\theta) = \exp(\theta)$, $R = \infty$], zero-truncated Poisson [$a(\theta) = \exp(\theta) - 1$, $R = \infty$], Hermite [$a(\theta) = \exp(\alpha\theta + \theta^2/2)$ with $\alpha > 0$ fixed, $R = \infty$], logarithmic series [$a(\theta) = -\log(1-\theta)$, $R = 1$], negative binomial [$a(\theta) = (1-\theta)^{-v}$ with $v > 0$ fixed, $R = 1$] and geometric [$a(\theta) = \theta(1-\theta)^{-1}$, $R = 1$].

The binomial distribution is the common example of the linear exponential law with finite support. If Y is distributed according to $B(N, p)$ a binomial distribution with number of trials N and success parameter p, the probability of the event $\{Y = k\}$ can be written

© Board of the Foundation of the Scandinavian Journal of Statistics 2005.

$$\binom{N}{k} p^k (1-p)^{N-k} = \binom{N}{k} \frac{\theta^k}{(1+\theta)^N} = \pi_{\Theta,k}, \qquad k = 0, 1, \ldots, N, \tag{3}$$

with $\theta = p/(1-p)$. Below, we shall use both parameterizations to define the binomial law. Note that in the case of finite $\mathbb{K}$ the radius R is infinity and the probabilities $\pi_{\theta,k}$, $k \in \mathbb{K}$ with $\theta = R$ can be defined by continuity. This means that in the binomial case we shall suppose $p \in [0, 1]$ or, equivalently, $\theta \in [0, \infty]$. Other examples of families of PSD with finite support can be obtained by (right-)truncation of the PSD with infinite $\mathbb{K}$. For the sake of brevity, in the case of finite $\mathbb{K}$ we shall reduce our attention to binomial laws, extensions to other linear exponential distributions with finite support being obvious.

A mixture of PSDs is a probability measure $\pi_Q = \{\pi_{Q,k}\}_{k \in \mathbb{K}}$ where

$$\pi_{Q,k} = \int_\Theta \pi_{\Theta,k} Q(\mathrm{d}\theta) = \int_\Theta \frac{a_k \theta^k}{a(\theta)} Q(\mathrm{d}\theta),$$

with Q the mixing distribution, that is a probability measure on Θ endowed with the Borel σ-field. Consider that the observations are distributed according to a mixture π_{Q_0}. The true mixing distribution Q_0 is unknown and it is supposed that $Q_0(\{0\}) < 1$. When $\mathbb{K}$ is finite the set Θ can be replaced by $[0, \infty]$ and in this case it is assumed that $Q_0((0,\infty)) > 0$.

In the case of infinite $\mathbb{K}$, we shall assume only that the support of Q_0 is contained in some compact $[0, M] \subset [0, R)$, but M is *unknown*. It is easy to see that, when R is finite, $Q_0([0, M]) = 1$, $M < R$ if and only if the moment generating function $\sum_{\mathbb{K}} \pi_{Q_0,k} e^{tk}$ is finite for some $t > 0$. When R is infinite, the compact support condition implies that the moment generating function is finite for any $t > 0$.

Several results on the identifiability of Q_0 have been proved (e.g. Sapatinas, 1995). Let us recall a simple identifiability result applicable when the support of the mixing distribution Q_0 is contained in a compact interval $[0, M] \subset [0, R)$. The following proposition was proved in the unpublished report of Milhaud & Mounime (1995). We reproduce the proof here for completeness.

Proposition 1

Assume that the support of the mixing distribution Q_0 is contained in a compact interval $[0, M] \subset [0, R)$. If $\sum_{k \in \mathbb{K},\, k>0} k^{-1} = \infty$, then Q_0 is identifiable.

Proof. Note that if $\sum_{k \in \mathbb{K},\, k>0} k^{-1} = \infty$, then $\mathbb{K}$ is necessarily an infinite set. First, we show that it suffices to prove that Q_0 is identifiable in the PSD mixture model with the mixing distributions concentrated on a subset of $[0, M]$. Indeed, assume that there exists Q_1 such that $\pi_{Q_1} = \pi_{Q_0}$ and $Q_1((M + a, R)) > 0$ for some $a > 0$. Then, there exists $C' > 0$ such that

$$\pi_{Q_1,k} \geq C' a_k (M+a)^k, \quad \text{for all } k \in \mathbb{K}.$$

Meanwhile, as $Q_0([0, M]) = 1$,

$$\pi_{Q_0,k} \leq C'' a_k M^k, \quad \text{for all } k \in \mathbb{K},$$

for some $C'' > 0$, and thus we contradict the equality between π_{Q_1} and π_{Q_0}. Consequently, $Q_1([0, M]) = 1$. Let $k_0 = \inf \mathbb{K}$. If $\pi_{Q_1} = \pi_{Q_0}$ with $Q_1([0, M]) = 1$, we write

$$\int_{[0,M]} \theta^{k-k_0} \tilde{Q}_0(\mathrm{d}\theta) = \int_{[0,M]} \theta^{k-k_0} \tilde{Q}_1(\mathrm{d}\theta), \quad \text{for all } k \in \mathbb{K},$$

© Board of the Foundation of the Scandinavian Journal of Statistics 2005.

where

$$\tilde{Q}_i(\mathrm{d}\theta) = \mu_{k_0}^{-1}\theta^{k_0}a(\theta)^{-1}Q_i(\mathrm{d}\theta), \quad i = 0, 1$$

and

$$\mu_{k_0} = \int_{[0,M]} \theta^{k_0}a(\theta)^{-1}Q_0(\mathrm{d}\theta) = \int_{[0,M]} \theta^{k_0}a(\theta)^{-1}Q_1(\mathrm{d}\theta).$$

By the Müntz–Szasz theorem (Rudin, 1987), the linear space spanned by the power functions $\{\theta^{k-k_0},\ k \in \mathbb{K}\}$ is dense in the space of continuous functions on [0, M] endowed with the supremum norm iff $\sum_{k\in\mathbb{K},k>k_0}(k-k_0)^{-1} = \infty$, that is iff $\sum_{k\in\mathbb{K},k>0}k^{-1} = \infty$. We deduce

$$\int_{[0,M]} f(\theta)\tilde{Q}_0(\mathrm{d}\theta) = \int_{[0,M]} f(\theta)\tilde{Q}_1(\mathrm{d}\theta)$$

for all continuous function on [0, M]. This implies $Q_1 = Q_0$.

The identifiability of the mixtures of binomials is studied in Lindsay (1995, Chapter 2). Let $B(N,\ p)$, $p \in [0,\ 1]$ denote the $\mathbb{R}^{N+1}$ vector of the binomial probabilities defined in (3). Basically, only the mixtures on the boundary of the convex hull $\mathbf{M} = co\{B(N,\ p),\ p \in [0,\ 1]\}$ are identifiable (see also Wood, 1999). Such mixtures are necessarily finite with at most $(N + 2)/2$ components. More precisely, the index of the mixing distribution should be at most N, where the index is the number of support points of the mixing distribution with the special rule that a support point equal to 0 or 1 be counted as 1/2.

3. ML estimation and rates of convergence

Let $X_1,\ldots, X_n \in \mathbb{K}$ be an i.i.d. sample distributed according to π_{Q_0}, a mixture of PSDs as defined in section 2. For simpler notation, we replace $\pi_{Q_0} = \{\pi_{Q_0,k}\}_{k\in\mathbb{K}}$ by $q_0 = \{q_{0,k}\}_{k\in\mathbb{K}}$. Define the log-likelihood function

$$l_n(Q) = \sum_{\mathbb{K}} \alpha_{n,k} \log \pi_{Q,k},$$

where $\alpha_n = \{\alpha_{n,k}\}_{k\in\mathbb{K}}$ is the vector of observed proportions. Let $\hat{Q}$ be the non-parametric maximum likelihood estimator (NPMLE), that is

$$l_n(\hat{Q}) = \sup_{Q} l_n(Q), \tag{4}$$

where the maximum is taken over all probability measures on Θ. For the models we consider, the NPMLE $\hat{Q}$ always exists because the log-likelihood l_n can be reconsidered as a strictly concave function defined on a compact and convex set of an Euclidean space (e.g. Lindsay, 1995, Chapter 5). Moreover, the corresponding estimator of the mixture, that is $\hat{\pi} = \pi_{\hat{Q}}$, is unique. Let us call $\hat{\pi}$ the NPMLE of the true mixture q_0.

When $\mathbb{K}$ is infinite, the support size, uniqueness and other finite sample properties of $\hat{Q}$ can be deduced using the same arguments as Simar (1976) and Lindsay (1995, Chapter 5).

In the case of non-parametric binomial mixtures, the uniqueness of $\hat{Q}$ depends on whether the empirical frequencies vector α_n is inside or outside the convex hull $\mathbf{M}$ defined above. If α_n does not belong to $\mathbf{M}$, the NPMLE $\hat{Q}$ is unique and quite easily computable, see Wood (1999, section 3). When α_n is inside $\mathbf{M}$ it means that the empirical distribution is a mixture of binomials and thus it maximizes the log-likelihood. In this case, in general, $\hat{Q}$ is no longer unique. Wood (1999) pointed out that finding a mixing distribution corresponding to $\alpha_n \in \mathbf{M}$ may be a delicate matter. However, he showed that, in practice, it is very likely that α_n will lie outside $\mathbf{M}$ for values of N greater than 10.

Let us point out that as the interest in this paper is to estimate the probabilities $q_{0,k}$, $k \in \mathbb{K}$, the fact that $\hat{Q}$ may not be unique is harmless as long as $\hat{\pi}$ is well-defined.

Concerning the asymptotic properties, using for instance the Hellinger distance and empirical process results, it can be proved that, almost surely, $\hat{\pi}_k \to q_{0,k}, k \in \mathbb{K}$ (e.g. Patilea, 2001). This result can be obtained without imposing either the uniqueness of $\hat{Q}$ or the identifiability of Q_0. Moreover, no particular assumption on $\mathbb{K}$ is required. If Q_0 is identifiable, almost surely, $\hat{Q} \to Q_0$ weakly.

In order to derive our asymptotic results we shall use the inner products and the χ^2-type norms introduced by Lambert & Tierney (1984). If $x \in \mathbb{R}^{\mathbb{K}}$ and π is a probability measure supported on $\mathbb{K}$, define $\|x\|_\pi = (\sum_{\mathbb{K}} x_k^2/\pi_k)^{1/2}$. Moreover, the inner product between x and y is defined by $\langle x,y\rangle_\pi = \sum_{\mathbb{K}} x_k y_k/\pi_k$ if $\|x\|_\pi, \|y\|_\pi < \infty$. In the following result and for the rest of this section we consider $\mathbb{K}$ unbounded.

Lemma 1

Assume that Q_0 *has a support included in a compact interval* $[0, M] \subset [0, R)$ *and that* Q_0 *is identifiable. Then, for any* $\varepsilon > 0$, *the quantities* $\|\alpha_n - q_0\|_{q_0}$, $\|\alpha_n - q_0\|_{\hat{\pi}}$ *and* $\|\hat{\pi} - q_0\|_{\hat{\pi}}$ *are of order* $o_{\mathbf{P}}(n^{-(1/2-\varepsilon)})$.

The rates of the three quantities above will be used to prove that, in some sense and under certain conditions, the difference between $\hat{\pi}$ and α_n is asymptotically negligible. The proof of the lemma above is identical to the one given in Lambert & Tierney (1984), proposition 3.1 (i), (ii) and (iv).

In addition to the three orders above, Lambert and Tierney also provided the order of $\|\hat{\pi} - q_0\|_{q_0}$, a key quantity for their proof of asymptotic normality. To prove that this last quantity is of order $o_{\mathbf{P}}(n^{-(1/2-\varepsilon)})$, they used the rate of $\|\hat{\pi} - q_0\|_{\hat{\pi}}$ and a suitable bound for $\hat{\pi}/q_0$. This bound is in fact a (non-random) vector $\{c_k\}_{k \in \mathbb{K}}$ such that, almost surely and for n large enough, $\hat{\pi}_k/q_{0,k} \le c_k$, $k \in \mathbb{K}$ and

$$\sum_{\mathbb{K}} q_{0,k} c_k^\delta < \infty \tag{5}$$

for some $\delta > 1/\varepsilon$. In the case of a PSD mixture with an infinite radius of convergence R, one can follow Lambert & Tierney (1984) assuming a known upper bound M for the support of Q_0 and restricting the maximum likelihood estimation to mixing distributions with the support in $[0, M]$. Then a bound for $\hat{\pi}/q_0$ that satisfies (5) is given by $c_k = C(M/m)^k$, $k \in \mathbb{K}$, with $0 < m < M$ such that $Q_0((m, M]) > 0$ and C is a positive constant.

Obtaining the rate of $\|\hat{\pi} - q_0\|_{q_0}$ becomes a delicate task when the radius of convergence of the power series is finite. Indeed, even if the support of Q_0 is included in a known compact subset $[0, M]$ of $[0, R)$ and the maximum likelihood estimation is restricted correspondingly, one needs to take m in the support of Q_0 arbitrarily close to M in order to ensure condition (5). In other words, when R is finite one has to know the upper limit of the support of Q_0 or at least one has to define, say $\hat{M}$, a suitable estimator of it. Milhaud & Mounime (1995) proposed such an estimator of M. In both situations, the maximum likelihood estimation has to be restricted to a compact interval subset of $[0, R)$, that is to $[0, M]$ or to $[0, \hat{M}]$.

Below, we show that the asymptotic normality of the estimated individual probabilities can be obtained without using the order of $\|\hat{\pi} - q_0\|_{q_0}$. Moreover, the estimates are obtained from unrestricted non-parametric maximum likelihood as defined in (4).

4. Asymptotic normality

As in Lambert & Tierney (1984), a key step is to show that under certain conditions on $x \in \mathbb{R}^{\mathbb{K}}$,

$$\sqrt{n}\langle \hat{\pi} - \alpha_n, x\rangle_{q_0} \to 0, \tag{6}$$

© Board of the Foundation of the Scandinavian Journal of Statistics 2005.

in probability. A classical central limit theorem then shows that

$$\sqrt{n}\langle\hat{\pi}-q_0,x\rangle_{q_0}\Rightarrow N(0,\sigma^2(x)),$$

where $\Rightarrow$ stands for the convergence in law and $\sigma^2(x)=\|x\|_{q_0}^2-\langle q_0,x\rangle_{q_0}^2$.

Note that the set of xs satisfying (6) is a linear subspace of $\mathbb{R}^{\mathbb{K}}$ and that q_0 belongs to this set. The main interest is to show that the unit vectors $e_i=\{e_{i,k}\}_{k\in\mathbb{K}}$ where $e_{i,k}=1$ if $k=i$ and 0 otherwise, satisfy (6). From this we deduce the same property for any vector x obtained as a linear combination of q_0 and the unit vectors. Other vectors with the desired property are also exhibited in this section.

Let $\bar{\pi}=(\hat{\pi}+q_0)/2$. The reason for showing (6) for the unit vectors is first to prove

$$\sqrt{n}\langle\hat{\pi}-\alpha_n,x\rangle_{\bar{\pi}}\to 0,$$

in probability, for a class of xs including the unit vectors and next to use the almost certain convergence of $\hat{\pi}$. First, consider the case of $\mathbb{K}$ unbounded. The case where $\mathbb{K}$ is finite will be examined at the end of the section.

Denote by $\mathcal{C}_1$ the set of x such that there exists a sequence $\{g_j\}$ of real-valued measurable bounded functions defined on Θ with (i) $\sup_{\Theta}|g_j(\theta)|\leq Kj^{\delta}$ for some $K,\delta>0$ and (ii) $||x-x(g_j)||_{q_0}=O(j^{-\beta})$ for some $\beta>0$, where

$$x_k(g_j)=\int_{\Theta}\frac{a_k\theta^k}{a(\theta)}g_j(\theta)Q_0(\mathrm{d}\theta). \tag{7}$$

Proposition 2

Assume Q_0 as in lemma 1. If $x\in\mathcal{C}_1$, then $\sqrt{n}\langle\hat{\pi}-\alpha_n,x\rangle_{\bar{\pi}}\to 0$, in probability.

Proof. Fix $x\in\mathcal{C}_1$ and let $\{x(g_j)\}$, K, δ and β be as in the definition of $\mathcal{C}_1$. Write $\sqrt{n}\langle\hat{\pi}-\alpha_n,\ x\rangle_{\bar{\pi}}=A_n+B_n$ with

$$A_n=\sqrt{n}\langle\hat{\pi}-\alpha_n,x-x(g_{j(n)})\rangle_{\bar{\pi}},\ B_n=\sqrt{n}\langle\hat{\pi}-\alpha_n,x(g_{j(n)})\rangle_{\bar{\pi}}$$

and $j(n)\sim n^{(1/2-\varepsilon)/\delta}$, $\varepsilon\in(0,1/2)$. Use the triangle inequality, Hölder's inequality and the fact that $\bar{\pi}_k/\sqrt{q_{0,k}}\geq\sqrt{\hat{\pi}_k}$ to deduce that

$$\begin{aligned}|A_n|&\leq\sqrt{n}\sum_{\mathbb{K}}\frac{|\hat{\pi}_k-q_{0,k}|+|q_{0,k}-\alpha_{n,k}|}{\bar{\pi}_k/\sqrt{q_{0,k}}}\frac{|x_k-x_k(g_{j(n)})|}{\sqrt{q_{0,k}}}\\&\leq\sqrt{n}\big(\|\hat{\pi}_k-q_{0,k}\|_{\hat{\pi}}+\|q_{0,k}-\alpha_{n,k}\|_{\hat{\pi}}\big)\|x-x(g_{j(n)})\|_{q_0}.\end{aligned}$$

As $\|x-x(g_{j(n)})\|_{q_0}=O(n^{-\beta(1/2-\varepsilon)/\delta})$ and in view of lemma 1, $A_n\to 0$ in probability. On the other hand, write $B_n=B_{1n}+B_{2n}$ with

$$\begin{aligned}|B_{1n}|&=\sqrt{n}\big|\langle\hat{\pi}-\alpha_n,x(g_{j(n)})\rangle_{\hat{\pi}}\big|\\&\leq\sqrt{n}\int_{\Theta}\left|\sum_{\mathbb{K}}\frac{\hat{\pi}_k-\alpha_{n,k}}{\hat{\pi}_k}\pi_{\theta,k}g_{j(n)}(\theta)\right|Q_0(\mathrm{d}\theta)\\&\leq K\sqrt{n}j(n)^{\delta}\int_{\Theta}\left|\sum_{\mathbb{K}}\frac{\hat{\pi}_k-\alpha_{n,k}}{\hat{\pi}_k}\pi_{\theta,k}\right|Q_0(\mathrm{d}\theta).\end{aligned}$$

By the gradient characterization of the non-parametric maximum likelihood (Lindsay, 1995, p.115), the absolute value in the last integral can be omitted and thus

© Board of the Foundation of the Scandinavian Journal of Statistics 2005.

$$|B_{1n}| \leq Kn^{1-\varepsilon} \sum_{\mathbb{K}} \frac{\hat{\pi}_k - \alpha_{n,k}}{\hat{\pi}_k} q_{0,k} = Kn^{1-\varepsilon} \sum_{\mathbb{K}} \frac{(\hat{\pi}_k - \alpha_{n,k})(q_{0,k} - \hat{\pi}_k)}{\hat{\pi}_k}.$$

Use Hölder inequality and lemma 1 and deduce $B_{1n} \to 0$, in probability. Finally,

$$\begin{aligned} |B_{2n}| &= \sqrt{n} \left| \langle \hat{\pi} - \alpha_n, x(g_{j(n)}) \rangle_{\overline{\pi}} - \langle \hat{\pi} - \alpha_n, x(g_{j(n)}) \rangle_{\hat{\pi}} \right| \\ &\leq K\sqrt{n} j(n)^{\delta} \sum_{\mathbb{K}} \frac{|\hat{\pi}_k - \alpha_{n,k}|}{\hat{\pi}_k} \frac{|\hat{\pi}_k - q_{0,k}|}{\hat{\pi}_k + q_{0,k}} q_{0,k} \\ &\leq Kn^{1-\varepsilon} \sum_{\mathbb{K}} \frac{(|\hat{\pi}_k - q_{0,k}| + |q_{0,k} - \alpha_{n,k}|)|\hat{\pi}_k - q_{0,k}|}{\hat{\pi}_k} \end{aligned}$$

and thus $B_{2n} \to 0$, in probability. Now, the proof is complete.

Lambert & Tierney (1984) showed that in the case of Poisson mixtures the unit vectors belong to class $\mathcal{C}_1$, provided that the true mixing distribution function grows faster than some power of θ in a neighbourhood of the origin. The same result remains true in the more general framework of PSD mixtures.

Assumption 1
There exist positive constants d, γ, ϵ such that $Q_0((\theta, \theta + \tau]) \geq d\tau^{\gamma}$ for all θ, $\tau \in (0,\epsilon)$.

Lemma 2
If assumption 1 holds, then for any $i \in \mathbb{K}$ the unit vector e_i belongs to $\mathcal{C}_1$.

The proof of this lemma is given in the appendix. Now, we can state the asymptotic normality for the estimated individual probabilities.

Corollary 1
Assume Q_0 as in lemma 1 and suppose that assumption 1 holds. Let $J = \{k_1,..., k_p\} \subset \mathbb{K}$. Then,

$$(\sqrt{n}(\hat{\pi}_{k_1} - q_{0,k_1}), \ldots, \sqrt{n}(\hat{\pi}_{k_p} - q_{0,k_p}), \sqrt{n}(\hat{\pi}_J^c - q_{0,J}^c)) \Rightarrow N(0, \Sigma),$$

where $\Sigma = diag(q_{0,k_1}, \ldots, q_{0,k_p}, q_{0,J}^c) - (q_{0,k_1}, \ldots, q_{0,k_p}, q_{0,J}^c)'(q_{0,k_1}, \ldots, q_{0,k_p}, q_{0,J}^c)$ and $\pi_J^c = 1 - \sum_{k \in J} \pi_k$ with $\pi = \hat{\pi}$ or q_0.

Proof. Since $\sqrt{n}\langle \hat{\pi} - \alpha_n, e_i \rangle_{q_0} = (\overline{\pi}_i / q_{0,i}) \sqrt{n} \langle \hat{\pi} - \alpha_n, e_i \rangle_{\overline{\pi}}$ and $\pi_i/q_{0,i} \to 1$, almost surely, we deduce that for any $i \in \mathbb{K}$ the unit vector e_i satisfies (6). The same remains true for the vector $x = q_0 - \sum_J q_{0,k} e_k$. Finally, use a classical central limit theorem.

Let us investigate further the class of vectors x satisfying (6). The following result, an extension of corollary 5.1 in Lambert & Tierney (1984), is proved in the appendix without using assumption 1.

Corollary 2
Assume Q_0 as in lemma 1. Let $\theta > 0$ such that $Q_0([\theta, R)) > 0$ and $Q_0((\theta - \tau, \theta + \tau])^{-1} = O(\tau^{-\delta})$ for some $\delta > 0$. Then,

$$\sqrt{n}\langle \hat{\pi} - \alpha_n, \pi_{\theta} \rangle_{q_0} \to 0, \tag{8}$$

© Board of the Foundation of the Scandinavian Journal of Statistics 2005.

and $\sqrt{n}\langle\hat{\pi}-\alpha_n, \pi'_\theta\rangle_{q_0} \to 0$, *in probability, where* $\pi'_\theta = \{\pi'_{\theta,k}\}_{k\in\mathbb{K}}$ *and* $\pi'_{\theta,k}$ *denotes the derivative of the map* $\theta \to \pi_{\theta,k}$. *Moreover, if the origin is an isolated point in the support of* Q_0, *then (8) holds also for* $\theta = 0$.

In the case where Q_0 has a finite support, in general, the conclusion of corollary 1 is no longer valid. Indeed, let x be a vector with $\sum_{\mathbb{K}} x_k = 1$, $x_k \geq 0$ and $x_k = 0$ except for a finite set of indices $J \subset \mathbb{K}$. The vector x plays the role of the observed proportions and it is supposed to be such that q_0 maximizes the corresponding likelihood, that is $\sum_J x_k \log \pi_{Q,k}$. The gradient characterization of the non-parametric maximum likelihood ensures that for any $\theta \in \Theta$, $\sum_J x_k \pi_{\theta,k}/q_{0,k} \leq 1$. As a consequence, for any sample size

$$\sum_{k\in J} \frac{x_k}{q_{0,k}} \left(\hat{\pi}_k - q_{0,k}\right) \leq 0, \tag{9}$$

which shows that the finite-dimensional vector $\left\{\sqrt{n}\left(\hat{\pi}_k - q_{0,k}\right)\right\}_{k\in J}$ cannot have a non-degenerate normal limit (see also Lambert & Tierney, 1984, p. 1398). However, the conclusion of corollary 2 is still true for the points in the support of Q_0.

4.1. The case of finite support $\mathbb{K}$

When the true distribution of the observations is a mixture of binomials in the interior of the convex hull **M**, the empirical distribution α_n almost certainly belongs to **M,** for n sufficiently large. In this case $\alpha_n = \hat{\pi}$ and the asymptotic normality becomes obvious.

When q_0 is on the boundary of **M** we can invoke the same type of arguments as were used to obtain (9). We deduce that the NPMLE $\hat{\pi}$ is no longer asymptotically equivalent to the empirical process. Nevertheless, the conclusion of corollary 2 remains valid. Indeed, as $\mathbb{K}$ is finite we have, almost certainly, $\hat{\pi}_k \to q_{0,k}$ uniformly in k, and this makes $n^{1/2}\|\alpha_n - q_0\|_{\hat{\pi}}$ and $n^{1/2}\|\hat{\pi} - q_0\|_{\hat{\pi}}$ bounded, in probability. The arguments used for proposition 2 again apply and thus corollary 2 can be proved. However, there will be at most N linearly independent vectors x satisfying (6).

5. Empirical evidence

5.1. Simulation experiments

We conducted a simulation experiment to compare the behaviour of $\hat{\pi}_J$ and $\alpha_{n,J}$. We also studied the non-parametric maximum likelihood (resp. empirical) estimator

$$\hat{\lambda}(k) = \frac{\hat{\pi}_k}{\sum_{j\geq k} \hat{\pi}_j} \qquad \left(\text{resp.} \quad \lambda_n(k) = \frac{\alpha_{n,k}}{\sum_{j\geq k} \alpha_{n,j}}\right)$$

of the hazard function $\lambda_0(k) = P(X = k)/P(X \geq k)$. The delta-method (e.g. van der Vaart, 1998) ensures that for any $k \in \mathbb{K}$,

$$\sqrt{n}(\lambda_n(k) - \lambda_0(k)) \Rightarrow N(0, V(k)) \quad \text{with} \quad V(k) = \frac{P(X \geq k+1)P(X = k)}{P(X \geq k)^3}.$$

Under the conditions of corollary 1, $\sqrt{n}(\hat{\lambda}(k) - \lambda_n(k)) \to 0$, in probability, and thus $\sqrt{n}(\hat{\lambda}(k) - \lambda_0(k))$ also converges in law to a normal $N(0,V(k))$.

Two types of compound families were used, namely Poisson and geometric ($\pi_{\theta,k} = \theta^{k-1}(1-\theta)$, $k = 1, 2,\ldots,\theta \in [0, 1)$). Given a mixture q_0 and a sample size n, we generated 1000 samples of size n from q_0. In each sample, we computed the NPMLE of the mixing distribution using the EM algorithm (e.g. Böhning, 2000). The algorithm started from the maximum

© Board of the Foundation of the Scandinavian Journal of Statistics 2005.

number of components for the NPMLE. The components with identical values or zero weights were collapsed. The NPMLE and the empirical proportions were used to compute probability mass and hazard function estimators.

For each set of 1000 estimates we used MINITAB (Minitab Inc., 2000) to build normal probability plots. Closeness to the fitting line (based on empirical mean and standard deviation) indicates good normal fit. We also report the Anderson–Darling goodness-of-fit statistic as a measure of how far the plot points lie from the fitted line. See Thode (2002) for a description of the Anderson–Darling test. MINITAB uses an adjusted Anderson–Darling statistic AD^*, in which points in the tails receive more weight. A smaller AD^* statistic indicates that the distribution provides a better fit to the data. Finally, note that a steeper fitted line indicates a smaller degree of variation for the data in the plot.

Two Poisson mixtures were considered. First, we studied NPMLE and the empirical distribution estimator (ED) of the probability $P(X = 0)$ when $Q_0 = 0.2\delta_0 + 0.8U[0, 4]$ and $n = 25$; δ_a denotes the probability measure concentrated at a and $U[0, 4]$ denotes the uniform distribution on [0, 4]. The true value of $P(X = 0)$ is 0.396. In Fig. 1, we present the probability plots for the 1000 values $\hat{\pi}_0$ and $\alpha_{n,0}$. The two types of estimates are quite close. The larger AD^* statistic for the empirical estimator is due to the lattice-valued nature of this estimator.

A significant difference between π_J and $\alpha_{n,J}$ appears when $P(X \in J)$ is small. Fig. 2 contains the probability plot for the estimates of $P(X = 6) = 0.022$ when the sample size is 100. We also considered $n = 25$ and $n = 50$ for which we obtained quite similar plots that are not presented here. As expected, the empirical proportion performs badly for small to moderate sample sizes. The law of the empirical estimator is far from normality. Meanwhile, the NPMLE of $P(X = 6)$ benefits from its smoother nature. It has lower variance (the fitted line is steeper) and its law is much closer to the normal approximation. It is noticeable that the same pattern remains valid even for larger sample sizes. For $n = 1000$ we found the AD^* statistic for NPMLE and empirical estimator of $P(X = 6)$ equal to 0.535 and 3.021, respectively (these values are also based on 1000 replications).

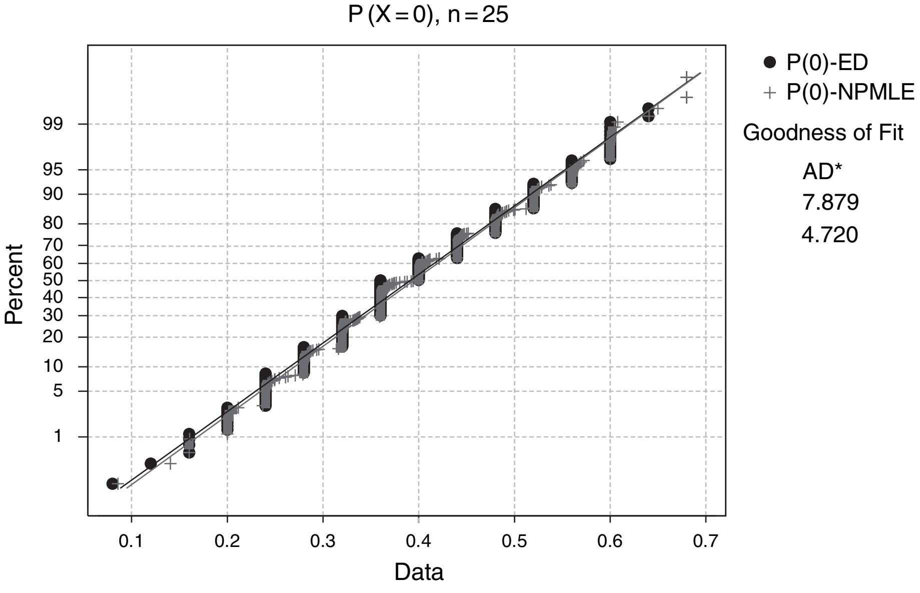

Fig. 1. NPMLE and empirical estimator of $P(X = 0)$: Poisson mixed with $Q_0 = 0.2\delta_0 + 0.8U[0, 4]$.

© Board of the Foundation of the Scandinavian Journal of Statistics 2005.

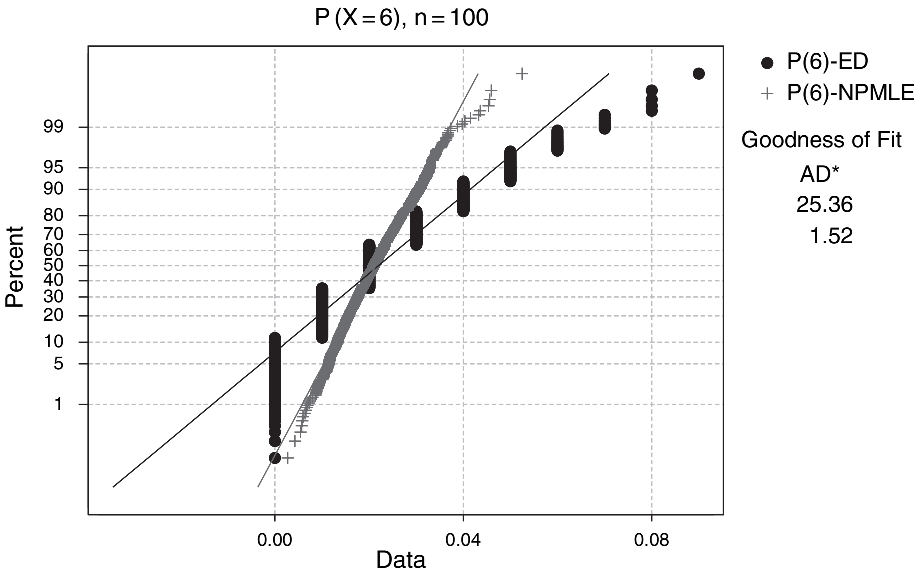

Fig. 2. NPMLE and empirical estimator of $P(X = 6)$: Poisson mixed with $Q_0 = 0.2\delta_0 + 0.8U[0, 4]$.

Fig. 3 shows the probability plots for NPMLE and empirical estimates of the hazard function value $\lambda(6) = 0.566$ when the sample size is 100. The rule $0/0 = 0$ was used when computing the empirical estimates. The plots indicate that empirical estimates are practically useless when the quantity of interest is the hazard function $\lambda(k)$ with k in the tails. The empirical estimates are concentrated at 0 and 1.

The second Poisson mixture considered has two components, more specifically $Q_0 = 0.85\delta_1 + 0.15\delta_4$. The quantities under study are $P(X = 6) = 0.016$ and $\lambda(6) = 0.491$ and n

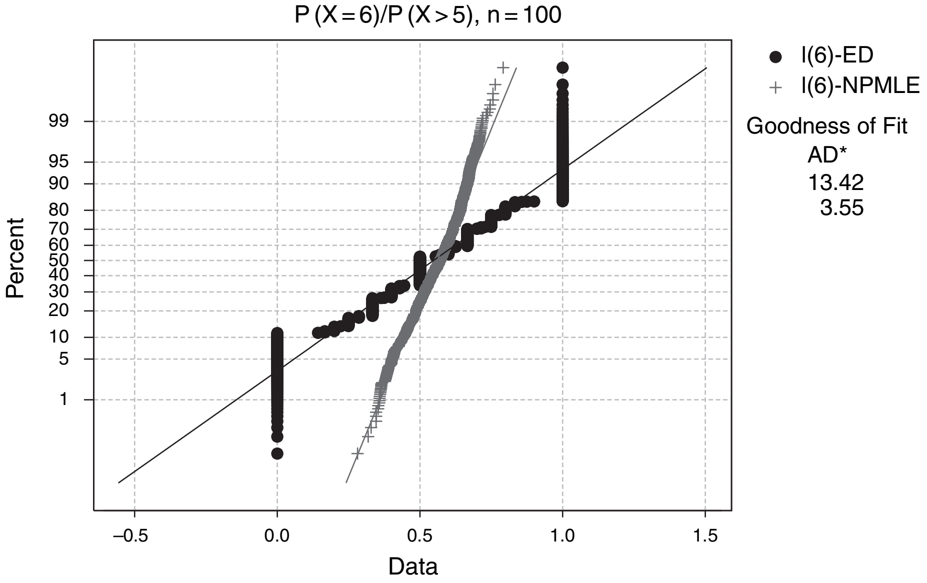

Fig. 3. NPMLE and empirical estimator of $\lambda(6)$: Poisson mixed with $Q_0 = 0.2\delta_0 + 0.8U[0, 4]$.

© Board of the Foundation of the Scandinavian Journal of Statistics 2005.

was taken equal to 25, 50 and 100. The results obtained are very similar to those obtained with the first Poisson mixture considered and therefore are not presented here, but are available from the authors. Corollary 2 does not apply for $\hat{\pi}_k$ when the true mixing distribution is discrete. However, the simulations indicate that the asymptotic normality of NPMLE for $P(X = k)$ and $\lambda(k)$ also holds in this case. The behaviour of NPMLE is like in the previous case, that is it performs much better than the empirical counterpart.

Finally, we computed the estimates of the tail probability $P(X \geq 6)$ of a geometric mixture defined by $Q_0 = U[0, 0.8]$ using samples of sizes $n = 25$ and $n = 100$. The true probability is 0.055. (The figures are not presented but they are available upon request.) Again, the NPMLE is characterized by less variability and its law is closer to the normal approximation (e.g. when $n = 100$ we obtained $AD^* = 9.523$ for the empirical proportions estimator and $AD^* = 2.079$ for the NPMLE).

5.2. Real data examples

Let us consider the Fabric Faults Data Sets analysed by Bissel (1972), see also Hinde (1982), McLachlan & Peel (2000). The data set is the number of faults in a bolt of fabric (see Table 1). The distribution of the counts has a remarkably long tail. We apply the non-parametric Poisson mixture model. We compute the NPMLE of the probabilities $P(X = k)$, $k = 0,1,..., 28$ and we compare them with the observed frequencies and the negative binomial (i.e. the parametric Poisson–Gamma mixture) fit.

Table 1. *Fabric faults data: Empirical, NPMLE and negative binomial estimates of $P(X = k)$.*

Counts	Observed frequency	NPMLE	Negative binomial
0	0	0.00655	0.01345
1	0.03125	0.01902	0.03236
2	0.03125	0.03332	0.05073
3	0.03125	0.05068	0.06534
4	0.125	0.07287	0.07503
5	0.03125	0.09574	0.07986
6	0.09375	0.11111	0.08053
7	0.125	0.1132	0.07799
8	0.09375	0.10217	0.07320
9	0.1875	0.08304	0.06700
10	0.0625	0.06209	0.06007
11	0	0.04404	0.05294
12	0	0.03093	0.04599
13	0	0.02265	0.03945
14	0.0625	0.01794	0.03348
15	0	0.01537	0.02815
16	0	0.01383	0.02347
17	0.0625	0.01266	0.01942
18	0	0.01157	0.01597
19	0	0.0105	0.01305
20	0	0.0095	0.01060
21	0	0.00859	0.00858
22	0	0.0078	0.00690
23	0.03125	0.00709	0.00554
24	0	0.00643	0.00442
25	0	0.00578	0.00352
26	0	0.00511	0.00279
27	0	0.00443	0.00221
28	0.03125	0.00375	0.00174

© Board of the Foundation of the Scandinavian Journal of Statistics 2005.

Owing to the same observed frequency, the empirical estimates of $P(X = k)$ for $k = 1, 2, 3, 5, 23$ and 28 are identical. The observed frequencies $\alpha_{n,k}$ are practically useless for estimating $P(X = k)$ for any unobserved value $k \leq 27$. Moreover, the empirical distribution yields quite bad estimators for the survivor function $P(X \geq k)$ and for the hazard function $\lambda(k) = P(X = k)/P(X \geq k)$. The NPMLE provide smoothed and more realistic estimates of all these quantities. The difference between the NPMLE and the negative binomial fit indicate a possible misspecification of the Poisson–Gamma model.

6. Discussion and conclusions

We studied mixtures of power series distributions using the non-parametric maximum likelihood approach. The asymptotic normality of the non-parametric maximum likelihood estimator of the probabilities $P(X \in J)$, J finite, was obtained. Our findings confirm the result of Lambert & Tierney (1984) for Poisson mixtures. The delta-method allows the extension of the asymptotic normality result to transformations of these probabilities, such as for instance the hazard function.

The NPMLE of $P(X \in J)$ has the same theoretical asymptotic behaviour as the empirical proportion $\alpha_{n,J}$; in particular it is efficient, without being a lattice-valued estimator. This allows the law of NPMLE of $P(X \in J)$ to quite quickly approach close to a normal law when sample size increases. Moreover, when used for quantities involving small $P(X \in J)$, NPMLE shows less variability than its empirical counterpart. In conclusion, due to its smooth nature, the NPMLE performs much better than its empirical competitor estimator in applications.

The two competing estimators $\hat{\pi}_J$ and $\alpha_{n,J}$ have the same asymptotic standard error $\sqrt{q_{0,J}(1 - q_{0,J})/n}$. In view of the theoretical results, two estimators can be used for this quantity:

$$\hat{\sigma}_1 = \sqrt{\frac{\hat{\pi}_J(1 - \hat{\pi}_J)}{n}} \qquad \text{or} \qquad \hat{\sigma}_2 = \sqrt{\frac{\alpha_{n,J}(1 - \alpha_{n,J})}{n}}.$$

Our simulations suggest that the variances of $\hat{\sigma}_1$ and $\hat{\sigma}_2$ could be quite different for small to moderate sample sizes. A consequence would be that Wald confidence intervals obtained from the asymptotic normality of $\hat{\pi}_J$ would have quite different performances when compared with Wald or other confidence intervals for binomial proportions based on $\alpha_{n,J}$. This issue will be analysed elsewhere.

Our theoretical results on the asymptotic normality do not include the quantities $\sqrt{n}(\hat{\pi}_k - q_{0,k}), k \in \mathbb{K}$ when q_0 the true law of the observations is a discrete mixture. This limitation is due to our approach based on the fact that, in some sense and under certain conditions, $\hat{\pi}$ and the empirical distribution are asymptotically equivalent. Equation (9) shows that this asymptotic equivalence no longer holds when the true mixture is discrete. However, we conjecture that for certain finite sets J including the singletons $J = \{k\}$, $k \in \mathbb{K}$, the asymptotic normality of $\sqrt{n}(\hat{\pi}_J - q_{0,J})$ also holds when the true law is a discrete mixture. Our simulations support this theory.

Acknowledgements

This research was partially undertaken while Valentin Patilea was at LEO, Université d'Orléans. The research of Dankmar Böhning was carried out with the support of the German Research Foundation (DFG). We would like to thank the Editor, the Associated Editor and the referees for their helpful comments.

References

Bissel, A. F. (1972). A negative binomial model with varying element sizes. *Biometrika* **59**, 435–441.

Böhning, D. (2000). *Computer-assisted analysis of mixtures and applications: meta-analysis, disease mapping and others*. Chapman & Hall/CRC London.

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

Carlin, B. R. & Louis, T. A. (1996). *Bayes and empirical Bayes methods for data analysis*. Chapman & Hall, London.

Gupta, P. L., Gupta, R. C. & Tripathi, R. C. (1997). On the monotonic properties of discrete failure rates. *J. Statist. Plann. Inference* **65**, 255–268.

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

Hinde, J. P. (1982). Compound Poisson regression models. In *GLIM 82* (ed. R. Gilchrist). Springer, New-York 109–121.

Kendall, M. & Stuart, A. (1991). *Advanced theory of statistics*, 2nd edn. Griffin, London.

Lambert, D. & Tierney, L. (1984). Asymptotic properties of the maximum likelihood estimates in mixed Poisson model. *Ann. Statist.* **12**, 1388–1399.

Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*, vol. 5. Institute of Mathematical Statistics and American Statistical Association, Hayward, California.

Mao, C. X. & Lindsay, B. G. (2002a). How many genes can be discovered if sequencing more ESTs? Technical Report, Pennsylvania State University, Philadelphia, PA.

Mao, C. X. & Lindsay, B. G. (2002b). A Poisson model for the coverage problem with a genomic application. *Biometrika* **89**, 669–681.

Martz, H. F., Kvam, P. H. & Abramson, L. R. (1996). Empirical Bayes estimation of the reliability of nuclear-power-plant emergency diesel generators. *Technometrics* **38**, 11–24.

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

Milhaud, X. & Mounime, S. (1995). A modified maximum likelihood estimator for infinite mixtures. Unpublished manuscript, Université de Toulouse 3.

Minitab Inc. (2000). *Minitab, Statistical Software, Release 13.30*. Minitab Inc., State College.

Patilea, V. (2001). Convex models, MLE and misspecification. *Ann. Statist.* **29**, 94–123.

Robbins, H. E. (1955). An empirical Bayes approach to statistics. *Proceedings of the Third Berkeley Symposium, Statistics and Probability*, Vol. 1. University of California Press, Berkeley 157–163.

Rudin, W. (1987). *Real and complex analysis*, 3rd edn. McGraw-Hill International Editions Singapore.

Sapatinas, T. (1995). Identifiability of mixtures of power-series distributions and related characterizations. *Ann. Inst. Statist. Math.* **47**, 447–459.

Shaked, M., Shanthikumar, G. & Valdez-Torres, J. (1994). Discrete probability orderings in reliability theory. *Statist. Sinica* **4**, 567–579.

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

Titterington, D. M., Smith, A. F. M. & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Wiley, New York.

Thode, H. C. (2002), *Testing for normality*. Marcel Dekker, New York.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press, Cambridge.

Wood, G. R. (1999). Binomial mixtures: geometric estimation of the mixing distribution. *Ann. Statist.* **27**, 1706–1721.

Received October 2003, in final form July 2004

Valentin Patilea, CREST-ENSAI, Campus de Ker-Lann, Rue Blaise Pascal – BP 37203, 35172 Bruz Cedex, France. E-mail: valentin.patilea@ensai.fr

Appendix

Proof of lemma 2.

The proof is an extension of the arguments given in theorem 4.2 of Lambert & Tierney (1984) (see also Milhaud & Mounime, 1995). Fix i and define

$$\lambda_{r,j} = \begin{cases} (a_i i!)^{-1} j^i \binom{i}{r} (-1)^{i-r} & r \leq i, \\ 0 & \text{otherwise} \end{cases}$$

and $I(r,j) = (r/j,\ r/j + j^{-(i+1)}]$, for $j = i,\ i+1,\ldots$ and $r = 0,1,\ldots,\ i$. Let

$$g_j(\theta) = \sum_{r=0}^{i} \lambda_{r,j} \mathbf{1}_{I(r,j)}(\theta) C(r,j), \quad \theta \in \Theta,$$

where $C(r,\ j)^{-1} = \int_{I(r,j)} a(\theta)^{-1} Q_0(\mathrm{d}\theta)$; $\mathbf{1}_A$ denotes the indicator function of the set A. The functions g_j are well defined for j sufficiently large. Use assumption 1 to show that g_j is bounded by a constant times $j^{i+\gamma(i+1)}$ as j grows to infinity. To calculate the order of $\|e_i - x(g_j)\|_{q_0}$ write

$$x_k(g_j) = \int_\Theta \frac{a_k \theta^k}{a(\theta)} g_j(\theta) Q_0(\mathrm{d}\theta) = a_k \sum_{r=0}^{i} \lambda_{r,j} \left(\frac{r}{j} + \xi_{r,j,k} \right)^k,$$

where $\xi_{r,j,k} \in (0, j^{-(i+1)}]$ is defined by the mean value theorem. The fact that $k \in \mathbb{K}$ is implicit throughout this proof. First, consider the case $k \leq i$ and bound $e_{i,k} - x_k(g_j)$. For some $\eta_{r,j,k} \in (0, j^{-(i+1)}]$,

$$x_k(g_j) = a_k \sum_{r=0}^{i} \lambda_{r,j} \frac{r^k}{j^k} + k a_k \sum_{r=0}^{i} \lambda_{r,j} \xi_{r,j,k} \left(\frac{r}{j} + \eta_{r,j,k} \right)^{k-1}.$$

Since $\xi_{r,j,k} \leq j^{-(i+1)}$, the second term of the right is of order $O(j^{-1})$. By lemma 4.2 of Lambert & Tierney (1984), the first term equals zero if $k < i$ and one if $k = i$. Now, it remains to bound $\sum_{k \geq i+1} x_k(g_j)^2 / q_{0,k}$. Deduce from the definition of $I(r,j)$ that

$$|x_k(g_j)| = \left| \sum_{r=0}^{i} \lambda_{r,j} \int_{I(r,j)} \frac{a_k \theta^k}{a(\theta)} C(r,j) Q_0(\mathrm{d}\theta) \right| \leq (i+1)^k j^{-k} a_k \sum_{r=0}^{i} |\lambda_{r,j}|.$$

Note also that $\sum_r |\lambda_{r,j}| \leq j^i 2^i / a_i i!$. Moreover, there exists $C,\ m > 0$ such that $q_{0,k} \geq C a_k m^k$, $k \in \mathbb{K}$. Therefore, for any $a \in (0,1)$

$$\sum_{k \geq i+1} \frac{x_k(g_j)^2}{q_{0,k}} \leq C(i) j^{-2a} \sum_{k \geq i+1} a_k s(j,k)^k,$$

where $s(j,\ k) = j^{-2(k-i-a)/k}(i+1)^2/m$ and $C(i)$ is a constant depending only on i. Since $\sup_{k \geq i+1} s(j,\ k) \to 0$ as $j \to \infty$, it follows that once j is large enough, the power series $\sum_{k \geq i+1} a_k s(j,\ k)^k$ is bounded by a constant. Thus, there exists $\beta > 0$ such that $\|e_i - x(g_j)\|_{q_0} = O(j^{-\beta})$.

Proof of corollary 2.

Fix $\theta > 0$ such that $Q_0([\theta,\ R)) > 0$ and $Q_0(A(\tau))^{-1} = O(\tau^{-\delta})$ for some $\delta > 0$, where $\mathbf{A}(\tau) = (\theta - \tau,\ \theta + \tau] \cap \mathit{supp} Q_0$. Define $g_\tau(\eta) = Q_0(A(\tau))^{-1} \mathbf{1}_{A(\tau)}(\eta)$ and consider the vectors $x(g_\tau) =$

© Board of the Foundation of the Scandinavian Journal of Statistics 2005.

$\{x_k(g_\tau)\}_{k\in\mathbb{K}}$, $\tau > 0$ as in (7). By the mean value theorem, for any $k \in \mathbb{K}$, there exists $\eta_k \in A(\tau)$ such that $x_k(g_\tau) = \pi_{\eta_k,k}$. Thus, again applying the mean value theorem we have

$$||\pi_\theta - x(g_\tau)||^2_{q_0} = \sum_{\mathbb{K}} \left(\pi_{\theta,k} - \pi_{\eta_k,k}\right)^2 q_{0,k}^{-1} \le \tau^2 \sum_{\mathbb{K}} \pi'^2_{\xi_k,k} q_{0,k}^{-1}, \tag{10}$$

for some $\xi_k \in A(\tau)$. Since $\pi'_{\eta,k} = a_k a(\eta)^{-1}\left\{k\eta^{k-1} - a'(\eta)\eta^k a(\eta)^{-1}\right\}$, there exists $C > 0$ such that $\|\pi'_{\eta,k}\| \le Ck\pi_{\eta,k}, \eta \in A(\tau)$, provided that τ is small enough. Moreover, the second order moments of π_η are uniformly bounded if η stays in a compact included in [0, R). Deduce that the sum on the right of equation (10) is finite. Consequently, $||\pi_\theta - x(g_\tau)||_{q_0} = O(\tau)$ as $\tau \to 0$ and thus $\pi_\theta \in \mathcal{C}_1$. If $\theta = 0$ is an isolated point in the support Q_0, then $\pi_0 = x(g_\tau) \in \mathcal{C}_1$ for sufficiently small τ. Apply proposition 2 and deduce that $\sqrt{n}\langle\hat{\pi} - \alpha_n, \pi_\theta\rangle_{\overline{\pi}} \to 0$, in probability. Finally, note that since $\pi_{\theta,k}Q_0([\theta,\ R)) \le q_{0,k}$, $k \in \mathbb{K}$,

$$\begin{aligned}\left|\langle\hat{\pi} - \alpha_n, \pi_\theta\rangle_{\overline{\pi}} - \langle\hat{\pi} - \alpha_n, \pi_\theta\rangle_{q_0}\right| &\le \sum_{\mathbb{K}} \frac{\left|\hat{\pi}_k - \alpha_{n,k}\right|\left|\hat{\pi}_k - q_{0,k}\right|}{\hat{\pi}_k + q_{0,k}} \frac{\pi_{\theta,k}}{q_{0,k}} \\ &\le Q_0([\theta, R))^{-1}(\|\hat{\pi} - q_0\|_{\hat{\pi}} + \|\alpha_n - q_0\|_{\hat{\pi}})\|\hat{\pi} - q_0\|_{\hat{\pi}}.\end{aligned}$$

For the convergence involving π'_θ deduce first that $\sqrt{n}\langle\hat{\pi} - \alpha_n, \pi_\theta\rangle_{\hat{\pi}} \to 0$, in probability. Next, proceed as in corollary 5.1 of Lambert & Tierney (1984). Assume that $\sqrt{n}\langle\hat{\pi} - \alpha_n, \pi'_\theta\rangle_{\hat{\pi}}$ does not converge to zero in probability for some $\theta > 0$ in the support of Q_0. Then, there exists a subsequence n_* for which almost surely $\sqrt{n_*}\sum\left(\hat{\pi}_k - \alpha_{n_*,k}\right)\pi'_{\theta,k}/\hat{\pi}_k$ is bounded away from zero but $\sqrt{n_*}\sum\left(\hat{\pi}_k - \alpha_{n_*,k}\right)\pi_{\theta,k}/\hat{\pi}_k$ converges to zero; here $\hat{\pi}$ is computed from n_* observations. Then, necessarily, for any large enough n_* there exists some η near θ such that $\langle\hat{\pi} - \alpha_{n_*}, \pi_\eta\rangle_{\hat{\pi}} < 0$. This contradicts the gradient characterization of the NPMLE.

© Board of the Foundation of the Scandinavian Journal of Statistics 2005.