Equivalence of Truncated Count Mixture Distributions and Mixtures of Truncated Count Distributions

Dankmar Böhning

School of Applied Statistics, School of Biological Sciences, University of Reading, Reading RG6 6FN, U.K.
email: d.a.w.bohning@reading.ac.uk

and

Ronny Kuhnert

Institute for Social Medicine, Epidemiology, and Health Economics, Charité Medical School Berlin,
Fabeckstr. 60-62, Haus 562, 14195 Berlin, Germany
email: ronny.kuhnert@charite.de

SUMMARY.   This article is about modeling count data with zero truncation. A parametric count density family is considered. The truncated mixture of densities from this family is different from the mixture of truncated densities from the same family. Whereas the former model is more natural to formulate and to interpret, the latter model is theoretically easier to treat. It is shown that for any mixing distribution leading to a truncated mixture, a (usually different) mixing distribution can be found so that the associated mixture of truncated densities equals the truncated mixture, and vice versa. This implies that the likelihood surfaces for both situations agree, and in this sense both models are equivalent. Zero-truncated count data models are used frequently in the capture-recapture setting to estimate population size, and it can be shown that the two Horvitz–Thompson estimators, associated with the two models, agree. In particular, it is possible to achieve strong results for mixtures of truncated Poisson densities, including reliable, global construction of the unique NPMLE (nonparametric maximum likelihood estimator) of the mixing distribution, implying a unique estimator for the population size. The benefit of these results lies in the fact that it is valid to work with the mixture of truncated count densities, which is less appealing for the practitioner but theoretically easier. Mixtures of truncated count densities form a convex linear model, for which a developed theory exists, including global maximum likelihood theory as well as algorithmic approaches. Once the problem has been solved in this class, it might readily be transformed back to the original problem by means of an explicitly given mapping. Applications of these ideas are given, particularly in the case of the truncated Poisson family.

KEY WORDS: Capture-recapture; Mixture of truncated count densities; Population size problem; Truncated mixture of count densities.

1. Introduction

Let $f(i, \lambda)$ denote a density for count variable value I, where I takes value $i \in \{0, 1, 2, \ldots\}$ and λ is an unknown parameter. The standard example is $f(i, \lambda) = Po(i, \lambda) = e^{-\lambda}\lambda^i/i!$, the conventional Poisson density. Suppose we are looking at zero truncation, for example, zeros are not observed. The associated truncated count density is $f_+(i, \lambda) = f(i, \lambda)/\{1 - f(0, \lambda)\}$ for $i = 1, 2, \ldots$. Zero truncation arises frequently in many practical situations. Two different categories are underlined:

- Zero counts cannot occur due to the observational model, for example, counts of occupants of passing cars. A study of pregnant women in mother-child health will lead to a nonzero count (gravida) of previous pregnancies including the present one. A computer-assisted telephone interviewing survey asking for the number of telephones in a household will necessarily have only nonzero counts in the sample. This situation is discussed and reviewed in Grogger and Carson (1991), Cameron and Trivedi (1998), and Winkelmann (2003).

- Zero-truncated count models are used in capture-recapture studies. From the individual capture history of a person, an animal, or an item, an attempt is made to predict the frequency of units missed by the sample. To illustrate, suppose that the police are keeping records in a particular community of the number of times a person has been identified with some deviant behavior. Clearly, those persons with deviant behavior who have never been identified will not be present in the database. One can use a truncated count model to predict this quantity. Other applications include the estimation of drug users in a community, and the suicide or terroristic potential

in a society. In economic studies, the number of illegal workers, for example, is of interest. This situation has been investigated mostly with emphasis on the Poisson distribution, for example, in Dahiya and Gross (1973), Blumenthal, Dahiya, and Gross (1978), Scollnik (1997), and van der Heijden et al. (2003). Also, Bunge and Fitzpatrick (1993) provide a review.

Mixtures of truncated count densities $f_+(i, \lambda_j)$ are given as

$$m_i = \sum_{j=1}^{k} p_j f_+(i, \lambda_j) \tag{1}$$

for $i = 1, 2, \ldots$, where the $p_j \geq 0$ and sum to 1. Alternatively, we might consider zero-truncated mixtures of untruncated densities $f(i, \lambda_j)$ given as

$$m'_i = \frac{\sum_{j=1}^{k} p'_j f(i, \lambda'_j)}{1 - \sum_{j=1}^{k} p'_j f(0, \lambda'_j)} \tag{2}$$

for $i = 1, 2, \ldots$, where again the $p'_j \geq 0$ and sum to 1. We call the discrete distribution $P = \begin{pmatrix} \lambda_1 & \cdots & \lambda_k \\ p_1 & \cdots & p_k \end{pmatrix}$ giving weight p_j to λ_j in (1) the *mixing distribution of size k*. Similarly for (2). For a general introduction to mixture models, see Titterington, Smith, and Makov (1985), Lindsay (1995), McLachlan and Krishnan (1997), McLachlan and Peel (2000), and Böhning (2000).

The mixture (1) of truncated count densities $\sum_{j=1}^{k} p_j f_+(i, \lambda_j)$ does *not* coincide with the truncated mixture of untruncated count densities (2) with $p'_j = p_j$ and $\lambda'_j = \lambda_j$. To illustrate, let $f(i, \lambda_j) = Po(i, \lambda_j)$ and the weights $p_j = p'_j \geq 0$ be arbitrary, but fixed. We choose $k = 2$, $p_1 = p_2 = p'_1 = p'_2 = 0.5$, and $\lambda_1 = \lambda'_1 = 1$, $\lambda_2 = \lambda'_2 = 4$. The two densities, corresponding to (1) and (2), are shown in Table 1.

In Table 1, in addition to the two mixture densities, the ratio of the truncated Poisson mixture to the mixture of trun-

Table 1
Comparing a mixture of truncated Poisson densities and a truncated mixture of untruncated Poisson densities; $k = 2$, all weights are 0.5, $\lambda_1 = 1$, $\lambda_2 = 4$

i	Truncated Poisson mixture	Mixture of truncated Poissons	Ratio
1	0.273355	0.328303	0.83263
2	0.204774	0.220124	0.93027
3	0.159053	0.148004	1.07465
4	0.130558	0.111630	1.16956
5	0.098747	0.082030	1.20380
6	0.064882	0.053474	1.21333
7	0.036940	0.030383	1.21579
8	0.018453	0.015170	1.21640
9	0.008199	0.006740	1.21656
10	0.003280	0.002696	1.21660
11	0.001193	0.000980	1.21661
12	0.000398	0.000327	1.21661
13	0.000122	0.000101	1.21661
14	0.000035	0.000029	1.21661
15	0.000009	0.000008	1.21661

cated Poisson densities is given in column 4. As it was pointed out in the editorial process, this ratio converges to a specific limit if I becomes large. It can be shown that this limit is given by

$$\frac{1 - (\exp(-\lambda_{\max}))}{1 - \{p_1 \exp(-\lambda_1) + \cdots + p_k \exp(-\lambda_k)\}} \geq 1$$

where $\lambda_{\max} = \max\{\lambda_1, \ldots, \lambda_k\}$.

Models (1) and (2) have different characteristics. Model (1) is *convex linear* in the mixing distribution, whereas model (2) is of practical working interest to the practitioner, in the sense that it is easier to interpret (the untruncated counts experience unobserved heterogeneity leading to the marginal distribution of a mixture, and truncation, consequently, leads to model (2)). Hence it is not surprising that it is typically model (2) that appears in the relevant literature (Norris and Pollock, 1996, 1998; Pledger, 2000; Link, 2003). The article will show that it is possible to work with model (1), which is considerably easier to treat, and to transfer results uniquely to model (2), and vice versa.

DEFINITION 1: Let the following sets be defined as:

$$M_k = \left\{ (m_1, m_2, m_3, \ldots)^T \mid m_i \right.$$
$$\left. = \sum_{j=1}^{k} p_j f_+(i, \lambda_j), \text{ for some } p_j, \lambda_j \right\}$$

$$M'_k = \left\{ (m'_1, m'_2, m'_3, \ldots)^T \mid m'_i \right.$$
$$\left. = \frac{\sum_{j=1}^{k} p'_j f(i, \lambda'_j)}{1 - \sum_{j=1}^{k} p'_j f(0, \lambda'_j)}, \text{ for some } p'_j, \lambda'_j \right\},$$

where λ_j, λ'_j, p_j, p'_j are all nonnegative, and $\sum_j p_j = \sum_j p'_j = 1$.

2. Results

Although the mixture of truncated count densities $\sum_{j=1}^{k} p_j f(i, \lambda_j)/\{1 - f(0, \lambda_j)\}$ does not coincide with the truncated mixture of untruncated count densities $\sum_{j=1}^{k} p_j f(i, \lambda_j)/\{1 - \sum_{j=1}^{k} p_j f(0, \lambda_j)\}$, the two sets M_k and M'_k are identical.

THEOREM 1: *The two sets of distributions are identical, that is, $M_k = M'_k$.*

Proof.

(a) We show $M'_k \subseteq M_k$. Let $m' \in M'_k$, in other words, there exists some $q_j, \lambda_j, j = 1, \ldots, k$ such that $m'_i = \sum_{j=1}^{k} q_j \times f(i, \lambda_j)/\{1 - \sum_{j=1}^{k} q_j f(0, \lambda_j)\}$. Define new weights

$$p_j = \frac{q_j\{1 - f(0, \lambda_j)\}}{\sum\limits_{h=1}^{k} q_h\{1 - f(0, \lambda_h)\}}.$$

Then,

$$\sum_{j=1}^{k} p_j f_+(i, \lambda_j) = \frac{\sum\limits_{j=1}^{k} q_j f(i, \lambda_j)}{1 - \sum\limits_{j=1}^{k} q_j f(0, \lambda_j)} = m_i'$$

and it follows that $m' \in M_k$.

(b) We show $M_k \subseteq M_k'$. Let $m \in M_k$, in other words, there exists some p_j, λ_j, $j = 1, \ldots, k$ such that $m_i = \sum_{j=1}^{k} p_j f_+(i, \lambda_j)$ for all $i = 1, 2, \ldots$. Define new weights

$$q_j = \frac{p_j/\{1 - f(0, \lambda_j)\}}{\sum\limits_{h=1}^{k} p_h/\{1 - f(0, \lambda_h)\}},$$

for $j = 1, \ldots, k$. Then,

$$\frac{\sum\limits_{j=1}^{k} q_j f(i, \lambda_j)}{1 - \sum\limits_{j=1}^{k} q_j f(0, \lambda_j)}$$

$$= \frac{\left[\sum\limits_{j=1}^{k} p_j f(i, \lambda_j)/\{1 - f(0, \lambda_j)\}\right] \Big/ \left[\sum\limits_{j=1}^{k} p_j/\{1 - f(0, \lambda_j)\}\right]}{1 - \left[\sum\limits_{j=1}^{k} p_j f(0, \lambda_j)/\{1 - f(0, \lambda_j)\}\right] \Big/ \left[\sum\limits_{j=1}^{k} p_j/\{1 - f(0, \lambda_j)\}\right]}$$

$$= \frac{\sum\limits_{j=1}^{k} p_j f(i, \lambda_j)/\{1 - f(0, \lambda_j)\}}{\sum\limits_{j} p_j/\{1 - f(0, \lambda_j)\} - \sum\limits_{j=1}^{k} p_j f(0, \lambda_j)/\{1 - f(0, \lambda_j)\}}$$

$$= \frac{\sum\limits_{j=1}^{k} p_j f_+(i, \lambda_j)}{\sum\limits_{j=1}^{k} p_j \left\{\frac{1}{1 - f(0, \lambda_j)} - \frac{f(0, \lambda_j)}{1 - f(0, \lambda_j)}\right\}}$$

$$= \sum_{j=1}^{k} p_j f_+(i, \lambda_j) = m_i,$$

as $p_1 + \cdots + p_k = 1$. It follows that $m \in M_k'$ and ends the proof.

Suppose frequencies $n_1, n_2, \ldots,$ for the different count values $1, 2, \ldots,$ have been observed in a sample. Let ν be the largest observed count. Then, the log likelihood is $l(P) = L(m) = \sum_{i=1}^{\nu} n_i \log(m_i)$ for model (1), where

$$m_i = m_i(P) = \sum_{j=1}^{k} p_j f(i, \lambda_j)/\{1 - f(0, \lambda_j)\}$$

and $l'(Q) = L(m') = \sum_{i=1}^{\nu} n_i \log(m_i')$ for model (2) with

$$m_i' = m_i'(Q) = \sum_{j=1}^{k} p_j' f(i, \lambda_j') \Big/ \left\{1 - \sum_{j=1}^{k} p_j' f(0, \lambda_j')\right\},$$

where P is giving weights $p_1, \ldots, p_k$ to $\lambda_1, \ldots, \lambda_k$ and Q is giving weights $p_1', \ldots, p_k'$ to $\lambda_1', \ldots, \lambda_k'$.

We mention a few consequences—a theorem and two corollaries—for the sake of brevity without proof.

THEOREM 2: *Let Ω_k denote the set of all discrete mixing distributions of size k. Then,*

$$\left\{(L(m), m)^T \mid m \in M_k\right\} = \left\{(L(m'), m')^T \mid m' \in M_k'\right\}$$

$$\left\{(l(P), P)^T \mid P \in \Omega_k\right\} = \left\{(l'(Q), Q)^T \mid Q \in \Omega_k\right\}.$$

COROLLARY 1: *Let $\hat{P}$ be the maximum likelihood estimate with respect to $l(P)$ for $P \in \Omega_k$ and let $\hat{Q}$ be the maximum likelihood estimate with respect to $l'(Q)$ for $Q \in \Omega_k$, if they exist. Then,*

$$l(\hat{P}) = L(m(\hat{P})) = L(m'(\hat{Q})) = l'(\hat{Q}).$$

A basic cornerstone of the analysis is the mapping

$$T : \Omega_k \ni P \to T(P) = Q \in \Omega_k,$$

which maps model (1) for any $P = \begin{pmatrix} \lambda_1 \cdots \lambda_k \\ p_1 \cdots p_k \end{pmatrix}$ to the equivalent model (2) with $Q = \begin{pmatrix} \lambda_1 \cdots \lambda_k \\ q_1 \cdots q_k \end{pmatrix}$, where $q_j = \frac{p_j/\{1 - f(0, \lambda_j)\}}{\sum_{h=1}^{} p_h/\{1 - f(0, \lambda_h)\}}$. Vice versa, the mapping

$$T' : \Omega_k \ni Q \to T'(Q) = P \in \Omega_k,$$

maps model (2) for any $Q = \begin{pmatrix} \lambda_1 \cdots \lambda_k \\ q_1 \cdots q_k \end{pmatrix}$ to the equivalent model (1) with $P = \begin{pmatrix} \lambda_1 \cdots \lambda_k \\ p_1 \cdots p_k \end{pmatrix}$, where $p_j = \frac{q_j\{1 - f(0, \lambda_j)\}}{\sum_{h=1}^{} q_h\{1 - f(0, \lambda_h)\}}$. In Figure 1, an illustration of the mappings T and T' is provided.
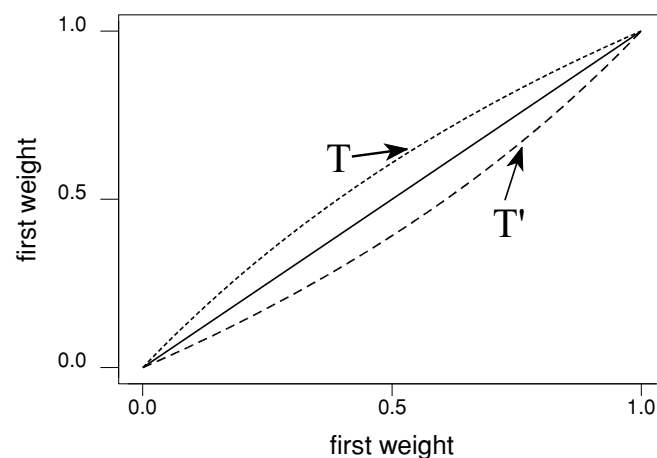
Figure 1. Illustration of the mappings T and T' for the Poisson case with $k = 2$, $\lambda_1 = 1$, $\lambda_2 = 4$.

With the help of T and T', the maximum likelihood estimates of the mixing distributions with k components can be equivalently provided.

COROLLARY 2: *Let $\hat{P}$ be the maximum likelihood estimate with respect to $l(P)$ for $P \in \Omega_k$ and let $\hat{Q}$ be the maximum likelihood estimate with respect to $l'(Q)$ for $Q \in \Omega_k$, if they exist. Then,*

$$T(\hat{P}) = \hat{Q} \quad and \quad T'(\hat{Q}) = \hat{P}.$$

3. Some Consequences

3.1 *Implications for the Horvitz–Thompson Estimator of the Population Size*

Assume that a sample of frequencies $n_1, n_2, \ldots$, of counts 1, 2, ..., has been observed. However, the frequency n_0 of zero counts is unobservable, and we wish to estimate it. Let $n = n_1 + n_2 + \cdots + n_\nu$ be the size of the observed sample, where ν is the largest count observed. Also, let $N = n_0 + n$ denote the unknown population size. Define the Horvitz–Thompson estimate of the population size as

$$\hat{N} = n \sum_{j=1}^{k} \frac{p_j}{1 - f(0, \lambda_j)}$$

based upon model (1) with $P = \begin{pmatrix} \lambda_1 & \cdots & \lambda_k \\ p_1 & \cdots & p_k \end{pmatrix}$ and

$$\hat{N}' = \frac{n}{1 - \sum_{j=1}^{k} p_j' f(0, \lambda_j')}$$

based upon model (2) with $P' = \begin{pmatrix} \lambda_1' & \cdots & \lambda_k' \\ p_1' & \cdots & p_k' \end{pmatrix}$.

THEOREM 3: *Let $P = \begin{pmatrix} \lambda_1 & \cdots & \lambda_k \\ p_1 & \cdots & p_k \end{pmatrix}$ and $Q = \begin{pmatrix} \lambda_1 & \cdots & \lambda_k \\ q_1 & \cdots & q_k \end{pmatrix}$ be connected by means of T and T' such that $T(P) = Q$ and $T'(Q) = P$. Then,*

$$\hat{N} = n \sum_{j=1}^{k} \frac{p_j}{1 - f(0, \lambda_j)} = \frac{n}{1 - \sum_{j=1}^{k} q_j f(0, \lambda_j)} = \hat{N}'.$$

Proof. Since $T'(Q) = P$ we have that $p_j = \frac{q_j \{1 - f(0, \lambda_j)\}}{\sum_h q_h \{1 - f(0, \lambda_h)\}}$ for all j. From here we have that

$$\sum_j p_j / \{1 - f(0, \lambda_j)\} = \frac{\sum_j q_j}{\sum_{h=1}^{\nu} q_h \{1 - f(0, \lambda_h)\}}$$

$$= \left\{ 1 - \sum_{h=1}^{\nu} q_h f(0, \lambda_h) \right\}^{-1},$$

and the result follows.

Theorem 3 implies that, if a P for model (1) is equivalent to a Q for model (2) (e.g., $T(P) = Q$ or $T'(Q) = P$), then the Horvitz–Thompson estimators of the population size are identical. In practice, P and Q need to be replaced by their max-

imum likelihood estimates and, given that $\hat{P}$ and $\hat{Q}$ are connected by means of T and T' (e.g., $T(\hat{P}) = \hat{Q}$ or $T'(\hat{Q}) = \hat{P}$), the associated population size estimators will agree.

3.2 *Some General Points for Mixtures of Truncated Densities*

From the analysis given so far it is clear that inference can be based upon the mixture model of truncated densities (1):

$$m_i = \sum_{j=1}^{k} p_j f_+(i, \lambda_j).$$

Now, assume that the nonparametric maximum likelihood estimator (NPMLE) $\hat{P}$ of the mixing distribution is available leading to $\hat{m}_i = \sum_{j=1}^{k} \hat{p}_j f_+(i, \hat{\lambda}_j)$. Then several questions arise with relation to *consistency*, *identifiability*, and *number of support points* for $\hat{P}$. van de Geer (2003) reviews asymptotic theory for maximum likelihood in nonparametric mixture models and shows consistency of the model space vector estimate $\hat{\mathbf{m}}$ under mild conditions for the mixture kernel, for example, $f_+(i, \lambda)$ in the case here. Consistency of $\hat{P}$ depends on the identifiability of the mixing distribution P. Lindsay and Roeder (1993) give strong results for discrete mixture models if the ν functions $f_+(1, \lambda), \ldots, f_+(\nu, \lambda)$ form a *Tchebycheff* system (Karlin and Studden, 1966) where again ν is the largest count observed. As a typical example, the truncated Poisson densities form a Tchebycheff system. In this setting Lindsay and Roeder (1993) establish that the *identifiable* mixing distributions are those with number of support points $\leq \nu/2$. Also, under these conditions the NPMLE of the mixing distribution is *unique*. If there are no restrictions on $f_+(i, \lambda)$, then it can be also shown as a consequence of the Theorem of Carathéodory (see Eggleston, 1966) that the NPMLE of the mixing distribution is discrete with number of support points bounded by ν (see also Lindsay, 1995).

In the following some results are discussed that do not require special conditions on $f_+(i, \lambda)$.

3.3 *Mixture Maximum Likelihood Theory*

The benefit of working with model (1) can be seen in the fact that an existing global maximization theory can be used. This was developed by various authors including Simar (1976), Laird (1978), Böhning (1982), Lindsay (1983), Leroux (1992), and Böhning (2000), among others. Let a sample of size n of zero-truncated counts be available and let $n_1, n_2, \ldots, n_\nu$ be their frequencies. Then, the log likelihood with respect to model (1)

$$l(P) = \sum_i n_i \log \left(\sum_{j=1}^{k} f_+(i, \lambda_j) p_j \right)$$

is a *concave functional* on the set of *all* discrete probability distributions (though it is not concave on Ω_k). This is the main reason for achieving the following global results.

An important analytical tool is the *gradient function* defined for any discrete distribution $P = \begin{pmatrix} \lambda_1 & \cdots & \lambda_k \\ p_1 & \cdots & p_k \end{pmatrix}$ as

$$d(\lambda, P) = \frac{1}{n} \sum_{i=1}^{\nu} n_i \frac{f_+(i, \lambda)}{f_+(i, P)},$$

where $f_+(i, P) = p_1 f_+(i, \lambda_1) + p_2 f_+(i, \lambda_2) + \cdots + p_k f_+(i, \lambda_k)$. With the help of the gradient function, the NPMLE can

be characterized. The general mixture maximum likelihood theorem (Böhning, 1982; Lindsay, 1983) states that for $\hat{P} = \left(\begin{smallmatrix} \hat{\lambda}_1 & \cdots & \hat{\lambda}_k \\ \hat{p}_1 & \cdots & \hat{p}_k \end{smallmatrix} \right)$

$$\hat{P} \text{ is NPMLE } \Leftrightarrow d(\lambda, \hat{P}) \leq 1 \quad \text{for all} \quad \lambda > 0. \qquad (3)$$

In addition, $d(\lambda, \hat{P}) = 1$ for $\lambda \in \{\hat{\lambda}_1, \ldots, \hat{\lambda}_k\}$, the set of all support points of $\hat{P}$. The benefit of the mixture maximum likelihood theorem for count densities like the truncated Poisson is even greater than for the untruncated Poisson family where other, simple diagnostic techniques like overdispersion tests are available (Böhning, 1994).

We will illustrate the mixture maximum likelihood theorem using a historic data set.

Example 1 (cholera epidemic in India). The numbers of cholera cases per house observed in a community in India of 223 houses were $n_1 = 32$, $n_2 = 16$, $n_3 = 6$, and $n_4 = 1$. These data have been used in numerous publications, but are frequently incompletely referenced. Lindsey (1995, p. 149) refers to Dahiya and Gross (1973) while Mao and Lindsay (2003) and Scollnik (1997) reference Blumenthal et al. (1978). Originally, the data were presented by McKendrick (1926) in his paper presentation to the Edinburgh Mathematical Society. According to Meng (1997), the ingenious work of McKendrick was ignored for a long time. Irwin (1963), as president of the Royal Statistical Society, wrote:

> McKendrick was in earlier life a lieutenant-colonel in the Indian Medical Service.... Though an amateur, he was a brilliant mathematician with a far greater insight than many professionals.

It can be assumed that McKendrick was confronted with the data of the cholera epidemic during the period of his service in India. It is interesting to note that there is also a number $n_0 = 168$ reported, the frequency of houses with *no* cholera cases. However, McKendrick knew that some unknown percentage of these houses was affected by the cholera epidemic, though no cases were observed in these houses. It should be recalled that cholera is a water-borne disease. In this case, the epidemic was caused by a specific, contaminated well, and houses supplied with drinking water from that contaminated well were exposed to developing the disease. This implies also that a certain number of houses were unexposed. Consequently, a simple Poisson distribution is not adequate, since at least the zero cell cannot be adequately described by the simple Poisson. To estimate the cholera-affected houses with no cases, two approaches are possible. One can either use a latent class approach leading to a zero-inflated Poisson distribution, or, alternatively, one can work only with the houses having at least one cholera case. (It can be shown that both approaches lead to the same result.) McKendrick chose to follow the second approach and developed a moment estimator for the parameter of the truncated Poisson distribution from which an estimator for the number of affected houses with no cases could then be derived. Dahiya and Gross (1973) investigated asymptotic properties of the McKendrick estimator. Irwin (1963) looked at maximum likelihood estimation for this case and proposed an iterative procedure which Meng (1997) used to illustrate the expectation-maximization (EM) algorithm. Blumenthal et al. (1978) investigated maximum likelihood estimation for
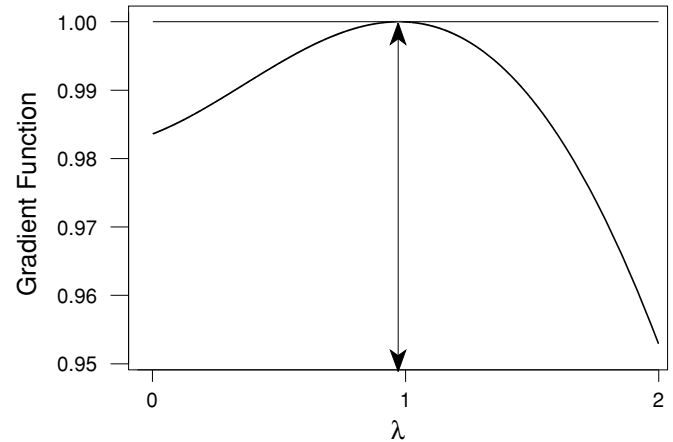
Figure 2. Gradient function for cholera data of McKendrick (1926) (arrow indicates position of MLE).

this case. At present, the McKendrick data are still used frequently without giving appropriate credit to the original contribution from McKendrick. Meng (1997) showed that McKendrick's moment estimator (and the improved maximum likelihood solution) provides an excellent fit to the data. This has been confirmed in Mao and Lindsay (2003). Here, we can also show that the nonparametric maximum likelihood estimator of the mixture of truncated Poisson distributions is a singular distribution giving all its mass to a single component with a mean of 0.972. A graph of the associated gradient function is given in Figure 2. The graph is below the bound 1, only becoming sharp at $\hat{\lambda} = 0.972$. No further investigation of heterogeneity is required. Indeed, as the discussion in Meng (1997) shows, the zero-truncated Poisson provides an excellent fit to the observed counts.

3.4 *Algorithms*

A variety of numerical algorithms exist to find the global maximum likelihood estimator, the NPMLE, if it exists. These include vertex direction methods and vertex exchange methods (Böhning, 2000) or intra-simplex direction methods (Lesperance and Kalbfleisch, 1992). However, it has become very popular to use the EM algorithm (Dempster, Laird, and Rubin, 1977) in connection with mixture models (McLachlan and Krishnan, 1997; McLachlan and Peel, 2000). The EM algorithm has the additional advantage of providing a maximum likelihood solution conditional upon the number of mixture components k though there is no guarantee for a nonlocal solution.

To proceed in the EM context we need the *complete data log likelihood*, which is given in this case as

$$\sum_{i=1}^{\nu} n_i \sum_{j=1}^{k} z_{ij} \log f_+(i, \lambda_j) + \sum_{i=1}^{\nu} n_i \sum_{j=1}^{k} z_{ij} \log p_j, \qquad (4)$$

where the unobserved covariate z_{ij} is 1 if i belongs to component j, and 0 otherwise. The EM algorithm replaces in the E-step the unobserved indicator variates z_{ij} by their expected values conditional upon the observed data and current values of λ_j, p_j, $j = 1, \ldots, k$ leading to

$$e_{ij} = E(z_{ij} \mid n_{ij}; p_j, \lambda_j, j = 1, \ldots, k) = \frac{f_+(i, \lambda_j)p_j}{\sum\limits_{j'=1}^{k} f_+(i, \lambda'_j)p'_j}. \quad (5)$$

In the M-step new values $\hat{\lambda}_1, \ldots, \hat{\lambda}_k, \hat{p}_1, \ldots, \hat{p}_k$ are found, which maximize the expected version of (4) leading to

$$\hat{p}_j = \frac{1}{n} \sum_{i=1}^{\nu} n_i e_{ij}, \quad \text{for } j = 1, \ldots, k \quad (6)$$

as new estimates for the weights. The new estimates $\hat{\lambda}_1, \ldots, \hat{\lambda}_k$ will depend on the form of $f_+(i, \lambda)$ and need to be found as solutions of

$$\sum_{i=1}^{m} n_i e_{ij} \frac{\partial}{\partial \lambda_j} \log f_+(i, \hat{\lambda}_j) = 0, \quad \text{for } j = 1, \ldots, k. \quad (7)$$

In the case of the Poisson, for example, $f_+(i, \lambda_j) = \frac{Po(i, \lambda_j)}{1 - \exp(-\lambda_j)}$, equation (7) takes the form

$$\hat{\lambda}_j = \frac{\sum\limits_{i=1}^{\nu} i\, n_i e_{ij}}{\sum\limits_{i=1}^{\nu} n_i e_{ij}}(1 - e^{-\hat{\lambda}_j}), \quad \text{for } j = 1, \ldots, k. \quad (8)$$

Note that (8) does not provide a closed-form solution for $\hat{\lambda}_j$, but rather suggests an iterative solution of the form $\hat{\lambda}_j^{\text{new}} = \frac{\sum_{i=1}^{\nu} i\, n_i e_{ij}}{\sum_{i=1}^{\nu} n_i e_{ij}}(1 - e^{-\hat{\lambda}_j^{\text{old}}})$, which needs to be iterated until convergence.

Example 2 (number of drug users in Bangkok).   In the second example, interest lies in estimating the number of illicit drug users in the Bangkok metropolitan area. Data stem from 61 private and public treatment centers in the region, which are permitted to treat drug addicts, and arose out of the surveillance system of the Office of the Narcotics Control Board (ONCB) of the Ministry of Public Health (Thailand) in the fourth quarter of the year 2001. Most of the drug users are heroin or methamphetamine users. Details of the study can be found in Böhning et al. (2004). For this application interest is solely focused on heroin users. The variable of interest is the number of occasions i a specific drug user visited one of the treatment institutions in the given time period. The data for heroin users are $n_1 = 2955$, $n_2 = 1186$, $n_3 = 803$, $n_4 = 611$, $n_5 = 416$, $n_6 = 338$, $n_7 = 278$, $n_8 = 180$, $n_9 = 125$, $n_{10} = 74$, $n_{11} = 38$, $n_{12} = 20$, $n_{13} = 14$, $n_{14} = 11$, $n_{15} = 4$, $n_{16} = 1$, $n_{17} = 3$, $n_{18} = 4$, $n_{19} = 1$. This means that 2955 drug users have visited one of the treatment centers exactly one time, 1186 have visited exactly two times, ... In total, there are $n = 7062$ heroin users, and interest is in finding an estimate of $N = n + n_0$.

Table 2 shows the results from the EM algorithm for mixtures of truncated Poisson distributions for various numbers of components k, starting from the homogeneous case $k = 1$. Note that the population size estimator is increasing with heterogeneity, illustrating a general monotonicity result for truncated count mixtures (Böhning and Schön, 2005). There is no further likelihood increase beyond four components, so

Table 2
Results of EM algorithm for various number of components k

k	Log likelihood	AIC	BIC	$\hat{N}$
1	-15462	-30927	-30934	7543
2	-13214	-26434	-26455	10226
3	-13134	-26279	-26313	13350
4	-13120	-26255	-26303	17278

that here the NPMLE is reached. It gives mass 0.3215, 0.4114, 0.2598, 0.0073 to Poisson components with respective parameters 0.1895, 2.0820, 5.8255, 12.1568. Figure 3 shows the gradient function for the NPMLE verifying that indeed the largest possible likelihood has been reached. Note that at the four support points of the mixing distribution the gradient function bound becomes sharp only here. To avoid oversmoothing compromise criteria that adjust for model complexity such as $AIC = 2 \times \log$ likelihood $- (2k - 1)2$, the Akaike information criterion, and $BIC = 2 \times \log$ likelihood $- (2k - 1)\log(n)$, the Bayesian information criterion, have been computed as well. The values of these statistics confirm that the full NPMLE is necessary for these data.

The gradient function can also be used to construct maximum likelihood estimates of the mixing distribution for fixed numbers of components, as suggested in Böhning (2003). The EM algorithm with gradient function update (EMGFU) starts with homogeneous ($k = 1$) maximum likelihood estimate $\hat{\lambda}$ and finds a new parameter point $\lambda_{\max}$ that maximizes the gradient function $d(\lambda, \hat{\lambda})$ in λ. Then, $\begin{pmatrix} \hat{\lambda} & \lambda_{\max} \\ \hat{\alpha} & (1-\hat{\alpha}) \end{pmatrix}$ is formed that serves as initial value for the EM algorithm with $k = 2$ components. The line maximizer $\hat{\alpha}$ is found to maximize the log likelihood $l((1 - \alpha)\delta_{\hat{\lambda}} + \alpha\delta_{\lambda_{\max}})$ in $\alpha \in [0, 1]$. Here, δ_x denotes the Dirac measure that puts all its mass at x. Suppose now the EM algorithm has iterated a maximum likelihood solution $\hat{P}_k$ for k components. Then, again, a new parameter point $\lambda_{\max}$ is found which maximizes the gradient function $d(\lambda, \hat{P}_k)$ in λ and form $\begin{pmatrix} \hat{\lambda}_1 & \cdots & \hat{\lambda}_k & \lambda_{\max} \\ (1-\hat{\alpha})\hat{p}_1 & \cdots & (1-\hat{\alpha})\hat{p}_k & \hat{\alpha} \end{pmatrix}$, which serves as initial value for the EM algorithm with $k + 1$ components.
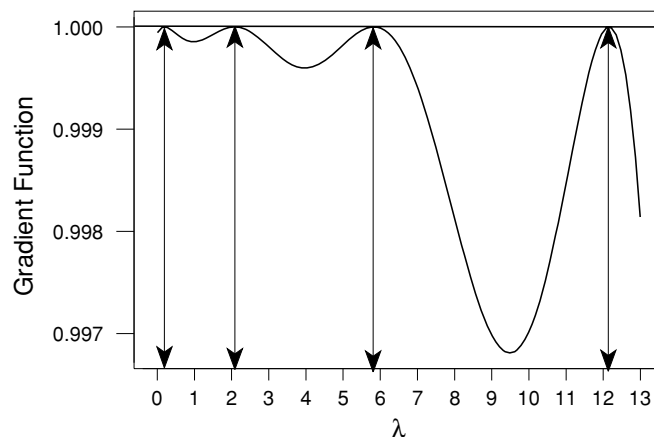

Figure 3.   Gradient function for heroin data (arrows indicate positions of support points of NPMLE).

Again, the line maximizer $\hat{\alpha}$ is found to maximize the log likelihood $l((1 - \alpha)\hat{P}_k + \alpha\delta_{\lambda_{\max}})$ in $\alpha \in [0, 1]$. This *forward selection* algorithm identifies the maximum likelihood estimates of the mixing distributions for all components k, starting with $k = 1$ and terminating at $k = \hat{k}$ defined by the NPMLE. This was used to construct the mixture maximum likelihoods in Table 2.

3.5 A Property of the NPMLE

For a truncated count variable Y let $f_+(y, Q) = \sum_{j=1}^{k} f_+(y, \lambda_j)q_j$ be the mixture of truncated densities where Q is the associated mixing distribution with k components. Also, assume that Λ is a discrete random variable with distribution Q, and expected value $E_Q(\Lambda)$. Then:

THEOREM 4:

$$E(Y) = E_Q\left\{\frac{e(\Lambda)}{1 - f(0, \Lambda)}\right\},$$

where $e(\lambda) = E(Y \,|\, \lambda) = \sum_{y=0}^{\infty} yf(y, \lambda)$, the conditional expectation of (untruncated) Y given λ.

Proof.

$$E(Y) = \sum_{y=1}^{\infty} yf_+(y, Q) = \sum_{y=1}^{\infty} y \sum_{j=1}^{k} f_+(y, \lambda_j)q_j$$

$$= \sum_{j=1}^{k} \left\{\sum_{y=1}^{\infty} yf_+(y, \lambda_j)\right\} q_j$$

$$= \sum_{j=1}^{k} \frac{\displaystyle\sum_{y=0}^{\infty} yf(y, \lambda_j)}{1 - f(0, \lambda_j)} q_j$$

$$= \sum_{j=1}^{k} \frac{e(\lambda_j)}{1 - f(0, \lambda_j)} q_j = E_Q\left\{\frac{e(\Lambda)}{1 - f(0, \Lambda)}\right\},$$

which ends the proof.

For the Poisson case the result is simply $E(Y) = E_Q\{\frac{\Lambda}{1 - \exp(-\Lambda)}\}$. The identity given in Theorem 4 also holds with population parameters replaced by their sample estimates. To see this we need the following lemma. We provide a simple form of this lemma although more general versions are possible. Let a sample of size n of zero-truncated counts be available with frequencies $n_1, n_2, \ldots, n_\nu$. For the untruncated case, there is also the frequency n_0 of zeros.

LEMMA 1: *Suppose the MLE of λ in the untruncated count density $f(i, \lambda)$ is the mean $\hat{\lambda} = \frac{\sum_{i=0}^{\nu} i\,n_i}{\sum_{i=0}^{\nu} n_i}$, then the MLE λ in the zero-truncated count density satisfies*

$$\hat{\lambda} = \frac{\displaystyle\sum_{i=1}^{\nu} i\,n_i}{\displaystyle\sum_{i=1}^{\nu} n_i}\{1 - f(0, \hat{\lambda})\}. \tag{9}$$

Proof. We consider the EM-steps for maximum likelihood estimation in zero-truncation models. We have in the E-step that $\hat{n}_0 = E(n_0 \,|\, \hat{\lambda}, n_1, \ldots, n_\nu) = \frac{f(0, \hat{\lambda})}{1 - f(0, \hat{\lambda})}n$. Using that in the M-step (untruncated situation) the maximum likelihood estimate is the sample mean we have

$$\hat{\lambda} = \frac{\displaystyle\sum_{i=0}^{\nu} i\,n_i}{n + \hat{n}_0} = \frac{\displaystyle\sum_{i=0}^{\nu} i\,n_i}{n + \dfrac{f(0, \hat{\lambda})}{1 - f(0, \hat{\lambda})}n} = \frac{\displaystyle\sum_{i=0}^{\nu} i\,n_i}{n}\{1 - f(0, \hat{\lambda})\},$$

and the result follows.

Now, we are able to give the following theorem, which is a modification of the (known) result for the untruncated case, namely, that the sample mean equals the mean of the mixing distribution (Proposition 10 in Lindsay, 1995; Karlis and Xekalaki, 1999).

THEOREM 5: *As in Lemma 1, it is assumed that the MLE of λ in the untruncated count density $f(i, \lambda)$ is the mean $\hat{\lambda} = \frac{\sum_{i=1}^{\nu} i\,n_i}{\sum_{i=1}^{\nu} n_i}$. Then*

$$\frac{\displaystyle\sum_{i=1}^{\nu} i\,n_i}{\displaystyle\sum_{i=1}^{\nu} n_i} = \sum_{j=1}^{k} \hat{p}_j \frac{\hat{\lambda}_j}{1 - f(0, \hat{\lambda}_j)}, \tag{10}$$

where $\hat{\lambda}_1, \ldots, \hat{\lambda}_k, \hat{p}_1, \ldots, \hat{p}_k$ are the maximum likelihood estimates for the mixing distribution for arbitrary, but fixed number of components k.

Proof. The complete expected data log likelihood for a mixture of truncated count densities with k components is given in (4). Taking partial derivatives, the necessity conditions (6) and (7) are deduced. Because of Lemma 1, equation (7) takes the form

$$\hat{\lambda}_j = \frac{\displaystyle\sum_{i=1}^{\nu} i\,n_i e_{ij}}{\displaystyle\sum_{i=1}^{\nu} n_i e_{ij}}\{1 - f(0, \hat{\lambda}_j)\}.$$

This implies that

$$\sum_{j=1}^{k} \hat{p}_j \frac{\hat{\lambda}_k}{1 - f(0, \hat{\lambda}_j)} = \sum_{j=1}^{k} \frac{1}{n} \sum_{i=1}^{\nu} n_i e_{ij} \frac{\displaystyle\sum_{i=1}^{\nu} i\,n_i e_{ij}}{\displaystyle\sum_{i=1}^{\nu} n_i e_{ij}}$$

$$= \frac{1}{n} \sum_{i=1}^{\nu} n_i \left(\sum_{j=1}^{k} e_{ij}\right) i = \frac{1}{n} \sum_{i=1}^{\nu} i\,n_i,$$

as e_{ij} is a probability distribution on j for fixed i (in fact, it is the posterior probability for count i to belong to component j), so that $\sum_{j=1}^{k} e_{ij} = 1$ for all i. This ends the proof.

A potential application and further value of Theorem 5 lies in the development of a stopping rule for the EM algorithm

for component numbers k lower than the NPMLE (for which a good stopping rule is available by means of the gradient function) or using it for reducing the labor in the EM algorithm. Theorem 5, for the truncated case, provides a modification of the result which, in the untruncated case, states that the sample mean equals the mean of the mixing distribution. This has been discussed in detail in Karlis and Xekalaki (1999) including using it as a diagnostic tool for checking convergence of the EM algorithm. It can be easily shown by means of the result from Karlis and Xekalaki that the maximum likelihood solution given by Hasselblad (1969) for the TIMES mortality data (consisting of the counts of death notices in the newspaper TIMES in the 3-year time period from 1910 to 1912 for women aged 80 and above) is *incorrect* (see also Titterington et al., 1985, p. 90) as is the solution given by Simar (1976) for the accident data of Thyrion (1960) (see also Böhning, 2000, p. 40).

To illustrate Theorem 5 for the data of the heroin users in example 2, it can be seen that the sample mean is $(2955 \times 1 + 1186 \times 2 + \cdots + 19 \times 1)/7062 = 2.9384 = 0.3215 \times 0.1895/(1 - \exp(-0.1895)) + \cdots + 0.0073 \times 12.1568/(1 - \exp(-12.1568))$, which corresponds to the numerical value of $\sum_{j=1}^{k} \hat{p}_j \frac{\hat{\lambda}_j}{1 - \exp(-\hat{\lambda}_j)}$.

4. Discussion

Recently there has been an increased interest in zero-truncated count models. These models can be used in social contexts such as illegal gun owners, illegal immigrants, illicit drug users, or car drivers without licenses to provide estimates of the size of the unobserved population. This article has shown that working with truncated mixtures of count densities and mixtures of truncated count densities is equivalent. This enables the researcher or practitioner to restrict oneself to the second, easier approach without losing generality.

Here, the results were formulated for a general parametric family of count densities $f(i, \lambda)$ with the Poisson as standard example in mind. As another example of a one-parametric family one could consider the binomial family, that is, a disease registry where a case is notified by y out of S sources where the sources notify independently with the same notification probability λ. Conditional upon S and the notifying probability λ the count of notifications Y will have a binomial distribution. Allowing nonparametric mixing here is appropriate to capture potential *source dependence* and *notification heterogeneity*. Most of the results given here will hold for the mixture of truncated binomials though results on identifiability and number of support points need modifications (see Lindsay and Roeder, 1993; Mao and Lindsay, 2002).

The algorithmic approach suggested here avoids the problem of choosing initial values, since it starts with the readily available maximum likelihood estimate for the homogeneous case ($k = 1$), and then increases the number of components in the mixing distribution by means of the gradient function, adding one component a time and updating the mixing distribution with the EM algorithm. As a by-product of this approach the occurrence of local maxima is reduced and the chances of reliably maximizing the likelihood in spaces with dimension lower than the one of the NPMLE are increased.

Finally, it is emphasized here that specifically the methods suggested for estimating population sizes need evaluation. This could be done by means of simulation or, even better, by means of existing data where true population size is known. This will be considered in future work.

References

Blumenthal, S., Dahiya, R. C., and Gross, A. J. (1978). Estimating the complete sample size from an incomplete Poisson sample. *Journal of the American Statistical Association* **73,** 182–187.

Böhning, D. (1982). Convergence of Simar's algorithm for finding the maximum likelihood estimate of a compound Poisson process. *Annals of Statistics* **10,** 1006–1008.

Böhning, D. (1994). A note on a test for Poisson overdispersion. *Biometrika* **81,** 418–419.

Böhning, D. (2000). *Computer-Assisted Analysis of Mixtures and Applications. Meta-Analysis, Disease Mapping and Others.* Boca Raton, Florida: Chapman & Hall/CRC.

Böhning, D. (2003). The EM algorithm with gradient function update for discrete mixtures with known (fixed) number of components. *Statistics and Computing* **13,** 257–265.

Böhning, D. and Schön, D. (2005). Nonparametric estimation of the population size based upon the counting distribution. *Journal of the Royal Statistical Society C* **54,** 721–731.

Böhning, D., Suppawattanabodee, B., Kusolvisitkul, W., and Viwatwongkasem, C. (2004). Estimating the number of drug users in Bangkok 2001: A capture-recapture approach using repeated entries in one list. *European Journal of Epidemiology* **19,** 1075–1083.

Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: A review. *Journal of the American Statistical Association* **88,** 364–373.

Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis of Count Data.* Cambridge, U.K.: Cambridge University Press.

Dahiya, R. C. and Gross, A. J. (1973). Estimating the zero class from a truncated Poisson sample. *Journal of the American Statistical Association* **68,** 731–733.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* **39,** 1–38.

Eggleston, H. G. (1966). *Convexity.* Cambridge, U.K.: Cambridge University Press.

Grogger, J. T. and Carson, R. T. (1991). Models for truncated counts. *Journal of Applied Econometrics* **6,** 225–238.

Hasselblad, V. (1969). Estimation of finite mixtures of distributions from the exponential family. *Journal of the American Statistical Association* **64,** 1459–1471.

Irwin, J. O. (1963). The place of mathematics in medical and biological statistics. *Journal of the Royal Statistical Society A* **126,** 1–45.

Karlin, S. and Studden, W. J. (1966). *Tschebysheff Systems: With Applications to Analysis and Statistics.* New York: Wiley.

Karlis, D. and Xekalaki, E. (1999). Improving the EM algorithm for mixtures. *Statistics and Computing* **9,** 303–307.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* **73,** 805–811.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *Annals of Statistics* **20,** 1350–1360.

Lesperance, M. L. and Kalbfleisch, J. D. (1992). An algorithm for computing the nonparametric MLE of a mixing distribution. *Journal of the American Statistical Association* **87,** 120–126.

Lindsay, B. G. (1983). The geometry of mixture likelihoods, part I: A general theory. *Annals of Statistics* **11,** 783–792.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry, and Applications.* NSF-CBMS Regional Conference Series in Probability and Statistics. Hayward, California: Institute of Statistical Mathematics.

Lindsay, B. G. and Roeder, K. (1993). Uniqueness and identifiability in nonparametric mixtures. *Canadian Journal of Statistics* **21,** 139–147.

Lindsey, J. K. (1995). *Modelling Frequency and Count Data.* Oxford: Oxford University Press.

Link, W. A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics* **59,** 1123–1130.

Mao, C. X. and Lindsay, B. G. (2002). Diagnostics for the homogeneity of inclusion probabilities in a Bernoulli census. *Sankhyā: The Indian Journal of Statistics, Series A* **64,** 626–639.

Mao, C. X. and Lindsay, B. G. (2003). Tests and diagnostics for heterogeneity in the species problem. *Computational Statistics and Data Analysis* **41,** 389–398.

McKendrick, A. G. (1926). Application of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society* **44,** 98–130.

McLachlan, G. and Krishnan, T. (1997). *The EM Algorithm and Extensions.* New York: Wiley.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* New York: Wiley.

Meng, X.-L. (1997). The EM algorithm and medical studies: A historical link. *Statistical Methods in Medical Research* **6,** 3–23.

Norris, J. L. III and Pollock, K. H. (1996). Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics* **52,** 639–649.

Norris, J. L. III and Pollock, K. H. (1998). Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species. *Environmental and Ecological Statistics* **5,** 391–402.

Pledger, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics* **56,** 434–442.

Scollnik, D. (1997). Inference concerning the size of the zero class from an incomplete Poisson sample. *Communication in Statistics—Theory and Methods* **26,** 221–236.

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

Thyrion, P. (1960). Contribution à l' étude du bonus pour non sinsitre en assurance automobile. *Astin Bulletin* **1,** 142–162.

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

van de Geer, S. (2003). Asymptotic theory for maximum likelihood in nonparametric mixture models. *Computational Statistics and Data Analysis* **41,** 453–464.

van der Heijden, P. G. M., Bustami, R., Cruyff, M. J. L. F., Engbersen, G., and van Houwelingen, H. C. (2003). Point and interval estimation of the population size using the truncated Poisson regression model. *Statistical Modelling—An International Journal* **3,** 305–322.

Winkelmann, R. (2003). *Econometric Analysis of Count Data.* Heidelberg: Springer-Verlag.