A Bagging-Based Correction for the Mixture Model Estimator of Population Size

Ronny Kuhnert[1], Victor J. Del Rio Vilas[2], James Gallagher[3], and **Dankmar Böhning[*,4]**

[1] Robert Koch-Institute, Division for Health of Children and Adolescents, Prevention Concepts, Seestr. 10, 13353 Berlin, Germany
[2] Centre for Epidemiology and Risk Analysis, Veterinary Laboratories Agency, New Haw, Addlestone, Surrey KT15 3NB, UK
[3] Statistical Services Centre, School of Biological Sciences, Harry Pitt Building, Whiteknights, Reading, RG6 6FN, UK
[4] Quantitative Biology and Applied Statistics, School of Biological Sciences, Harry Pitt Building, Whiteknights, Reading, RG6 6FN, UK

Received 10 February 2008, revised 4 July 2008, accepted 23 July 2008

Summary

Estimation of a population size by means of capture-recapture techniques is an important problem occurring in many areas of life and social sciences. We consider the frequencies of frequencies situation, where a count variable is used to summarize how often a unit has been identified in the target population of interest. The distribution of this count variable is zero-truncated since zero identifications do not occur in the sample. As an application we consider the surveillance of scrapie in Great Britain. In this case study holdings with scrapie that are not identified (zero counts) do not enter the surveillance database. The count variable of interest is the number of scrapie cases per holding. For count distributions a common model is the Poisson distribution and, to adjust for potential heterogeneity, a discrete mixture of Poisson distributions is used. Mixtures of Poissons usually provide an excellent fit as will be demonstrated in the application of interest. However, as it has been recently demonstrated, mixtures also suffer under the so-called boundary problem, resulting in overestimation of population size. It is suggested here to select the mixture model on the basis of the Bayesian Information Criterion. This strategy is further refined by employing a bagging procedure leading to a series of estimates of population size. Using the median of this series, highly influential size estimates are avoided. In limited simulation studies it is shown that the procedure leads to estimates with remarkable small bias.

Key words: Bagging; Bootstrap; Boundary Problem; Nonparametric Mixture Model; Population Size Estimator; Zero-truncation.

1 Introduction

The size N of some population is often unknown and requires determination. It is assumed that the identifying mechanism leaves a number of the members of the population of interest undetected. In the biological sciences this is often a wildlife population (and the identifying mechanism is an animal trap) whereas in the life or social sciences this population might be a group of people who are difficult to sample such as illicit drug users (with identifying mechanism a hospital register) or car drivers without a license (with identifying mechanism a police data base). Suppose that a specific mechanism identifies some, say n, but not all units of a population of size N. Furthermore, assume that identifica-

* Corresponding author: e-mail: d.a.w.bohning@reading.ac.uk, Phone: +44(0) 11 83 78 62 11, Fax: +44(0) 11 378 8032

tion occurs independently for each population unit with probability $1-p_0$. This stochastic situation can be described by tuples of size N

$$(\delta_1, \delta_2, \ldots, \delta_N)$$

where $\delta_i = 1$ indicates that the i-th unit is identified (and observed) and $\delta_i = 0$ otherwise (and the unit remains unobserved). Each of these tuples occur with probability $(1-p_0)^{\sum_{i=1}^{N}\delta_i} p_0^{N-\sum_{i=1}^{N}\delta_i}$. We are interested in the probability that exactly n units are identified. Since there are $\binom{N}{n}$ tuples $(\delta_1, \delta_2, \ldots, \delta_N)$ with $\sum_{i=1}^{N}\delta_i = n$ the probability of observing exactly n units is a simple *binomial probability*:

$$\binom{N}{n}(1-p_0)^n p_0^{N-n}. \tag{1}$$

Then, the maximum likelihood estimator of N is the well-known Horvitz-Thompson estimator (Horvitz and Thompson, 1952) of the population size given as

$$\hat{N} = \frac{n}{(1-p_0)}. \tag{2}$$

Note that (1) can be viewed as a likelihood function in N which is maximized for N being the integer part of (2) (Lindsay and Roeder, 1987). However, p_0 will be known only in exceptional cases and usually an estimate of p_0 will be required for practical use. In addition, some modeling for p_0 will be required. We will address this in the following section. For a more general introduction into the capture-recapture methodology see Bunge and Fitzpatrick (1993).

2 Capture-Recapture Studies Leading to Frequencies of Repeated Identifications

The mechanism that identifies units with probability $1-p_0$ can be quite general. It might be that several sources identify the units leading to a log-linear modeling approach for the estimation of p_0 and N (Bishop et al., 1975). Another common method for deriving an estimator of p_0 is based upon counting repeated identifications of the same unit by the same mechanism over a given time span. This is usually referred to as capture-recapture data in the form of frequencies of frequencies. For example, in a capture-recapture study repeated occurrences of dolphins are counted by some mechanism or the number of times a patient receives treatment for a certain disease at a medical facility may be counted. We will denote by f_0, f_1, $f_2, \ldots, f_m$ the frequency of those units identified exactly $0, 1, 2, \ldots, m$ times where m is the largest occurring count. Also, we will denote with p_0, p_1, $p_2, \ldots, p_m$ the probability of exactly $0, 1, 2, \ldots. m$ identifications. Clearly, f_0 is unobserved and target of the inference. We have that $n = f_1 + f_2 + \cdots + f_m$ and $N = n + f_0$.

Example 1 Sheep are kept in holdings in Great Britain and the occurrence of scrapie is monitored by the Compulsory Scrapie Flocks Scheme. This was established in 2004 and summarizes three surveillance sources. The frequency distribution of the *scrapie count within each holding* for the year 2005 is presented in Table 1. See also Del Rio Vilas and Böhning (2008) for further details. Pre-

Table 1 Scrapie data for Great Britain 2005 (Del Rio Vilas and Böhning 2008).

Number of scrapie cases	0	1	2	3	4	5	6	7	8	n
Frequency of holdings	–	84	15	7	5	2	1	2	2	118

viously, the under-ascertainment adjusted prevalence has been estimated via anonymous postal surveys (Hoinville et al., 2000; Sivam et al., 2003). More recently, multiple-list capture-recapture methods were applied to estimate the number of scrapie-affected holdings not detected by any of the surveillance streams in place (Del Rio Vilas et al., 2005). In the following we attempt to develop a methodology for providing an adjustment for disease undercount.

Example 2 To illustrate the frequencies of frequencies situation we look at the following capture-recapture data: Oremus (2005) tried to estimate the size of a small community of spinner dolphins which are resident around the island of Moorea (near Tahiti). In 2002, over an interval of 8 months, skin samples were randomly taken and 12 microsatellite loci were genotyped which makes mis-matching of dolphins very unlikely. $f_1 = 42$ dolphins were sampled only once, $f_2 = 7$ dolphins were sampled exactly twice and $f_3 = 2$ dolphins were sampled exactly three times. This leads to $n = 51$ different dolphins that were observed in the experiment. For more details see Böhning (2008).

3 Mixture Modelling and the Boundary Problem

The problem of modelling the probability p_j for observing count j arises, where $j = 0, 1, 2, \ldots$. The Poisson density $Po(j; \lambda) = \exp(-\lambda)\,\lambda^j/j!$ does not often provide enough flexibility to give an adequate fit. Mixture models (Norris and Pollock, 1996, 1998; Pledger, 2000; Mao and Lindsay, 2002, 2003) are more flexible, and we consider a discrete mixture of Poisson distributions of the form

$$f(j; Q_k) = \sum_{\ell=1}^{k} Po(j; \lambda_\ell)\, q_\ell\,, \tag{3}$$

where the mixing distribution $Q_k = \begin{pmatrix} \lambda_1 & \ldots & \lambda_k \\ q_1 & \ldots & q_k \end{pmatrix}$ is giving weight $q_\ell \geq 0$ to parameters λ_ℓ for $\ell = 1, \ldots, k$, and k is the number of components (Poisson densities) in the mixture. Note that $q_1 + \ldots + q_k = 1$. For a general introduction into mixture models see the books of Titterington et al. (1985) and Lindsay (1995). Böhning and Schön (2005) discuss maximum likelihood estimation for a given number of components k. Likelihood analysis focuses on the zero-truncated mixture log-likelihood

$$\log L(Q_k) = \sum_{j=1}^{m} f_j \log\left[f(j, Q_k)\right] - n \log\left[1 - f(0, Q_k)\right]. \tag{4}$$

Equivalently, a log-likelihood based upon mixtures of zero-truncated Poisson distributions could be considered (Böhning and Kuhnert, 2006). In this situation the log-likelihood can be maximized in the set of all discrete probability distributions, leading to the *nonparametric maximum likelihood estimate* (NPMLE). For more details see the appendix. For the surveillance data on scrapie (see Table 1) the NPMLE corresponds to $k = 3$ components. Details of the likelihood analysis for $k = 1$ to 3 components, including the associated maximised log-likelihood, are presented in Table 2. Column 1 in Table 2 contains the number of components in the mixture model (3) and it can be seen that with $k = 3$ the nonparametric maximum likelihood is achieved (see column 2). The differences in the likelihoods

Table 2 Mixture likelihood analysis for the scrapie data of Table 1.

Number of components k	$\log L(\hat{Q}_k)$	BIC	$\hat{f}_0$	$\hat{N}$
1	−155.9	313.9	52	170
2	−126.9	260.0	274	392
3 (NPMLE)	−126.4	263.2	1117	1235

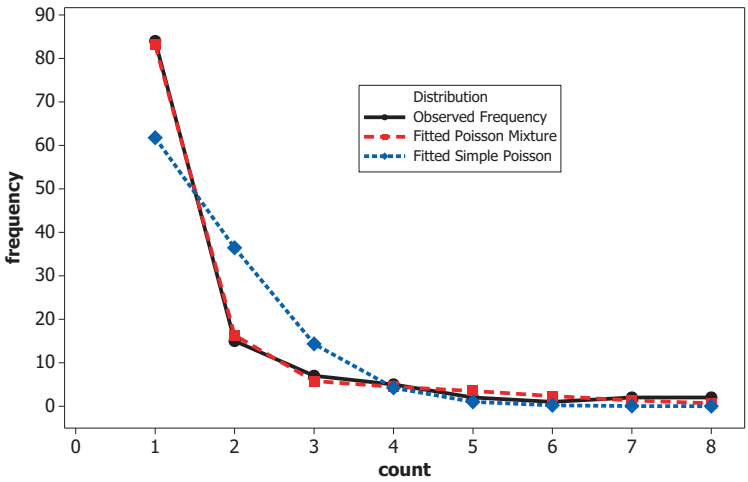

Figure 1 Frequency distribution of observed counts and count distributions fitted by simple Poisson and mixture of two Poisson distributions; data are from Example 1.

for models with $k = 2$ and $k = 3$ components are minor, which is clearly evident when the *Bayesian Information Criterion (BIC)* is considered:

$$\text{BIC} = -2 \log L\left(\hat{\boldsymbol{Q}}_k\right) + (2k - 1) \log(n)\,. \tag{5}$$

The BIC penalizes the log-likelihood with the number of parameters $(2k - 1)$ multiplied by the log-sample size and works well as model selection criterion in mixture model settings as it does not suffer under likelihood irregularities that are typical for mixture models (Chen et al., 2001; McLachlan and Peel, 2000). Models are selected on the basis of small BIC-values: the smaller the BIC-value, the better the model. According to the analysis provided in Table 2, the model of choice is the two-component model. Having identified the model and the associated parameter estimates we can estimate the probability for a zero count p_0 as

$$\hat{p}_0 = \sum_{\ell=1}^{k} Po(0; \hat{\lambda}_\ell)\, \hat{q}_\ell = \sum_{\ell=1}^{k} \exp\left(-\hat{\lambda}_\ell\right) \hat{q}_\ell \tag{6}$$

so that $\hat{N} = n/(1 - \hat{p}_0)$ and $\hat{f}_0 = \hat{N} - n$. Results for the scrapie data are presented in Table 2. As can be seen in Fig. 1, the two-component mixture model provides a good fit to the observed frequencies whereas the simple Poisson is clearly not adequate. It is however crucial that a selection criterion is employed, such as the BIC, which penalizes for oversmoothing the data. As can also be seen from the analysis in Table 2, the NPMLE although providing an excellent fit, also carries the risk of overestimation of the population size, potentially drastically. Not only do practitioners consider the estimate of $\hat{N} = 1235$ as unrealistically high, but well-established alternative nonparametric population size estimators such as Chao's lower bound estimator $\hat{N}_{\text{C}} = n + f_1^2/(2f_2) = 353$ (Chao, 1987) and Zelterman's robust estimator $\hat{N}_{\text{Z}} = n/[1 - \exp(-2f_2/f_1)] = 393$ (Zelterman, 1988) are either close or in the vicinity of the estimate $\hat{N} = 392$ from the BIC-selected mixture model with $k = 2$ components (and not close to the NPMLE given with $k = 3$ components).

Example 2 (continued). To illustrate the overestimation in an extreme case we look at the spinner dolphins data again. Table 3 provides the results of the mixture model based likelihood analysis. Evidently, the one-component model is the right choice leading to an estimate of 153 for the population size. The alternative estimators of Chao with 177 and Zelterman 180 are close. The nonparametric

Table 3 Mixture likelihood analysis for the spinner dolphin data of Example 2.

Number of components k	$\log L(\hat{Q}_k)$	BIC	$\hat{f}_0$	$\hat{N}$
1	−29.1	59.2	102	153
2 (NPMLE)	−28.9	61.1	111,678	111,729

maximum likelihood estimate is provided for $k = 2$ components leading to a *spurious* estimate of 111, 729. This example shows that results based on the non-parametric mixture maximum likelihood estimator should be considered with great caution.

The problem becomes even more apparent if the following simulation experiment is considered. $N = 100$ counts are sampled from a two-component Poisson mixture giving equal weights to component means $\lambda_1 = 1$ and $\lambda_2 = 3$. The frequencies $f_0, f_1, \ldots, f_m$ are constructed and f_0 is ignored. The experiment is repeated 100 times and $\hat{N}$ is computed on the basis of the NPMLE for each of the 100 samples. The results are presented as an individual value plot on the left-hand-side of Fig. 2. It can be seen that there are a large number of values overestimating the true value dramatically. The mean estimate was 31,856 and the largest estimate of N was 263,163 which indicates how useless the non-parametric maximum likelihood approach can become in this situation. However, it is also clear from Fig. 2 that there are numerous reasonable good estimated values in the vicinity of the true population size. Hence the approach suffers from the occurrence of many influential points and it appears wise to choose a robust estimate of population size. In the simulation study this is easy to accomplish by looking at the median which is 109 for the scenario above. Although there is a tremendous improvement in reducing *overestimation bias*, a slight overestimation bias appears to persist. A different strategy follows the BIC-selected modelling approach. For the scenario above we find the majority of estimated population sizes (estimated on the basis of the BIC-selected mixture model) in the vicinity of the true population size (see the right-hand-side of Fig. 2). However, even here large influential population size estimates occur and lead to a mean estimate of 7,369, vastly overestimating the true size of 100. The median with a value of 92 does better and we will suggest a general estimation strategy based on the median of BIC-selected population size estimates.

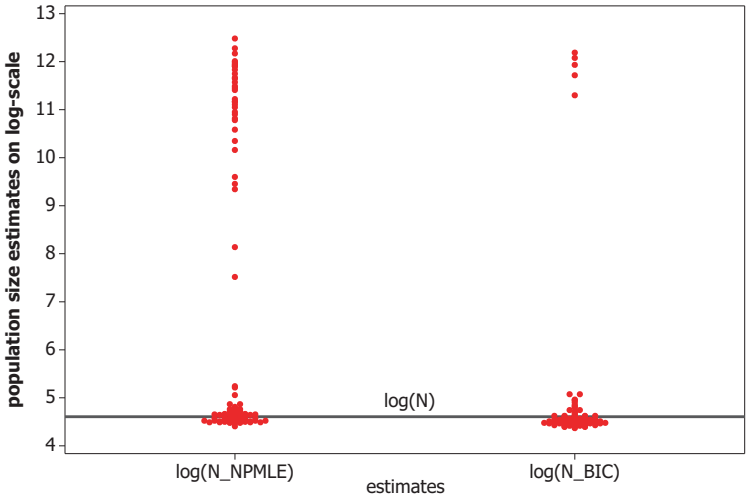

Figure 2 Distribution of estimates based upon the NPMLE and the BIC-selected mixture model.

The reason for this (potentially severe) overestimation bias has been debated for some time. One of the reasons, potentially the central reason, is the so-called *boundary problem* which describes the fact that for untruncated Poisson mixture models $Nf(0;\hat{Q}) \geq f_0$ and $E(f(0;\hat{Q})) \geq p_0$. The result is due to Harris (1991) and covered in more generality by Wang and Lindsay (2008). See also the associated part in the appendix.

4 A Median-Correction for the Mixture Model Estimator of Population Size

If repeated samples are available it is a simple task to consider a diversity of location estimators such as the mean or, as was suggested here, the median. Unfortunately, only one sample of frequencies of counts $f_1, f_2, \ldots, f_m$ is available in practice, leading to only one estimate $\hat{N}$ of N. Further, this estimate does not carry any salient characteristic which let practitioners decide if it is a "trustworthy" or spurious observation.

We will utilize available techniques for improving *unstable* estimators, in particular, we will consider *bagging* (**b**ootstrap **agg**regat**ing**) as suggested by Breiman (1996) which we will modify for our purposes. According to Bühlmann and Yu (2002) (who also provide a theoretical foundation for the method) *bagging* consists of three steps: (i) B bootstrap samples are constructed from the original sample, (ii) for each of the B bootstrap samples the estimator of interest is computed, and (iii) the expected value of the estimator with respect to the bootstrap distribution is computed. The last step is usually implemented by replacing the expected value with the sample mean of the B estimators generated by each of the B bootstrap samples. This mean is called the *bagged* estimator. The benefit of the method lies in the fact that the variance of the bagged estimator is reduced in comparison to the original estimator. As Bühlmann and Yu (2002) point out, this gain can be drastic if the original estimator is "unstable", e.g. small changes in the data can lead to large changes in the estimator. For our purposes, we need to modify step iii) since large values for the population size estimator will affect the mean which is known to be sensitive to large values. Instead, we will use the median as a summary measure for the bootstrap estimates.

Hence, we suggest the following nonparametric bootstrap procedure:

Nonparametric Bootstrap

1. Sample n counts $Y_1^*, \ldots, Y_n^*$ from a multinomial distribution with size parameter n and category probability parameters f_j/n.
2. Construct from this sample the frequencies $f_1^*, \ldots, f_m^*$.
3. Construct $\hat{N}^*$ using the BIC-selected mixture model.

Suppose the above nonparametric bootstrap has been used to generate B samples of size n each and there are now B population size estimates $\hat{N}_1^*, \ldots, \hat{N}_B^*$ available where each of these has been determined on the basis of a BIC-selected mixture model. Then we are able to determine a variety of measures including the median. We define the *median-adjusted bootstrap estimator* or *bagged estimator* of population size as

$$\hat{N}_M = \text{median}\{\hat{N}_1^*, \ldots, \hat{N}_B^*\}. \tag{7}$$

Note that the above bootstrap algorithm is different from the one suggested in van der Heijden et al. (2003) in that it samples from the observed distribution $f_1/n, f_2/n, \ldots, f_m/n$ in contrast to sampling from $\hat{f}_0/\hat{N}, f_1/\hat{N}, f_2/\hat{N}, \ldots, f_m/\hat{N}$ which is required for obtaining an estimate of the variance of $\hat{N}$.

Example 1 (continued) We apply the bootstrap algorithm to the scrapie surveillance data presented in Example 1. A bootstrap sample was generated on the basis of the observed $f_1, \ldots, f_8$. All mixture models up to the nonparametric maximum likelihood estimate were generated and the one with the

Table 4 Population size estimators for the scrapie surveillance data based upon the median-corrected mixture model estimator and the alternative estimators of Chao and Zelterman (computation is based upon the bootstrap distribution).

Location measure	BIC-mixture model	Chao	Zelterman
median	396	364	401
mean	3717	374	417

best BIC-value selected. Also the alternative estimators of Chao and Zelterman were computed for comparison. This process was repeated 100 times. Table 4 shows the median and mean estimates for this bootstrap distribution. Whereas there are only minor differences between the two for the estimators of Chao and Zelterman, there is a considerable difference for the mixture-model based estimator which shows the importance of the median-correction in this case. In fact, the large value for the mean was caused by just one single value in the bootstrap distribution of the mixture-model based estimator of population size (see Fig. 3).

5 A Simulation Study

Although the median-adjusted bootstrap estimator of population size behaved well in the examples, there is no guarantee that this will generally be the case. For a more systematic approach we undertook the following simulation experiment. Counts we generated from a two-component mixture of Poisson distributions with equal weights attached to the component means λ_1 and λ_2. The population size to be estimated was $N = 1000$. For each simulated data set $f_0, f_1, f_2, \ldots, f_m$ was determined with $f_0 + f_1 + f_2 + \ldots + f_m = N$. Then, f_0 was ignored and the zero-truncated frequencies $f_1, f_2, \ldots, f_m$ were used to determine the BIC-selected mixture model. This model was used to estimate N on the basis of (6). In addition, for each of the simulation samples a bootstrap sample $f_1^*, \ldots, f_m^*$ was generated and the population size estimated on the basis of the BIC-selected mixture model. This was done for each of $B = 30$ bootstrap samples. Then, the median of the bootstrap estimates of the population size determined.

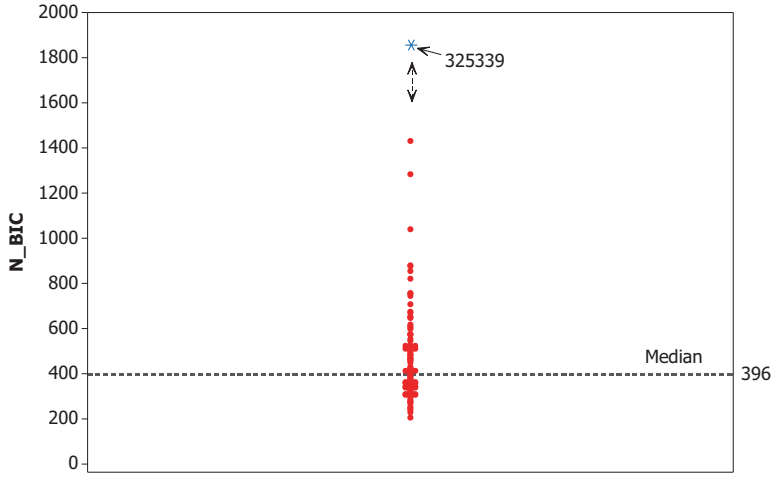

Figure 3 Distribution of estimates based upon the NPMLE and the BIC-selected mixture model.

Table 5 Median of population size estimators of Zelterman, Chao and the median-adjusted bootstrap mixture model (median absolute deviations) for 8 mixture populations of the form $0.5Po(1) + 0.5Po(\lambda_2)$; medians are based on 100 replications and median-adjusted mixture model estimator is based upon 30 bootstrap replications, true population size is $N = 1000$.

λ_2	Zelterman	Chao	Median – adjusted mixture model	
			BIC-selected	NPMLE
1	992 (39)	991 (35)	1000 (25)	1031 (48)
2	987 (27)	975 (20)	967 (30)	1098 (145)
3	996 (25)	952 (14)	1010 (37)	1083 (102)
4	1034 (32)	952 (21)	1003 (39)	1042 (57)
5	1096 (39)	960 (16)	997 (24)	1038 (53)
6	1157 (48)	972 (20)	999 (24)	1032 (47)
7	1197 (58)	982 (21)	998 (25)	1032 (46)
8	1239 (55)	985 (19)	1001 (21)	1027 (37)

Table 6 Median of population size estimators of Zelterman, Chao and the median-adjusted bootstrap mixture model (median absolute deviations) for 8 mixture populations of the form $0.5\,Po\,(1) + 0.5\,Po\,(\lambda_2)$; medians are based on 100 replications and median-adjusted mixture model estimator is based upon 30 bootstrap replications, true population size is $N = 500$.

λ_2	Zelterman	Chao	BIC-selected
1	495 (29)	496 (23)	503 (18)
2	493 (20)	486 (15)	479 (17)
3	499 (18)	478 (13)	506 (27)
4	518 (22)	475 (10)	503 (21)
5	557 (28)	486 (15)	503 (22)
6	591 (31)	488 (13)	508 (13)
7	615 (40)	496 (15)	504 (13)
8	622 (46)	494 (19)	506 (12)

The results are found in Tables 5–7, for $N = 1000$, $N = 500$, $N = 100$, respectively. Following Wang and Lindsay (2008) to adjust for the skewness of the distribution, we present the median of all population size estimators as well as the associated median absolute errors. The median-adjusted bootstrap mixture model estimator with BIC-selection of the number of components has consistently smaller median bias than Chao's estimator and Zelterman's estimator, which we have included in the computations for comparison. In Table 5 we have also included the bagged NPMLE. It is overestimating and, hence, BIC-selection of mixture models appears appropriate. The overestimation bias of the bagged NPMLE is similar and relatively constant over the different populations and it is less than Zelterman's bias which becomes large with increasing values of the second component. Note that the median absolute error of the median-adjusted mixture model estimators is comparable or even lower than Zelterman's estimator. However, it is clear that Chao's lower bound estimator has the better

Table 7 Median of population size estimators of Zelterman, Chao and the median-adjusted bootstrap mixture model (median absolute deviations) for 8 mixture populations of the form $0.5\,Po\,(1) + 0.5\,Po\,(\lambda_2)$; medians are based on 100 replications and median-adjusted mixture model estimator is based upon 30 bootstrap replications, true population size is $N = 100$.

λ_2	Zelterman	Chao	BIC-selected
1	100 (11)	99 (9)	100 (8)
2	99 (9)	98 (8)	96 (6)
3	101 (10)	96 (7)	92 (6)
4	109 (12)	98 (6)	103 (13)
5	112 (16)	97 (8)	104 (14)
6	117 (17)	98 (7)	102 (8)
7	124 (20)	101 (8)	103 (9)
8	128 (22)	101 (9)	101 (9)

standard error. Several other mixture model constellations had been considered (in particular with more than two components) with similar results. Hence, the results are not reported here. The simulation results have been developed with a stand-alone program which is available from the authors upon request.

6 Discussion

Discrete mixture models offer a wide and flexible modelling framework to cope with heterogeneity in the parameters representing capture-recapture probabilities. They are potentially the most suitable models for fitting recapture counts – as has been demonstrated by many authors (Mao and Lindsay, 2002, 2003; Norris and Pollock, 1996, 1998; Pledger, 2000). However, when discrete mixture models are to predict unobserved zero counts and hence the population size, overestimation bias may be severe. The bias, due to the boundary problem, may not be obvious to a practitioner implementing an estimation procedure to obtain the nonparametric maximum likelihood estimator. Consequently, adjustments to the model selection procedure are required with respect to the number of components in the mixture model. We have suggested using the BIC criterion for model selection. Other criteria are possible, but these have not been considered here since the BIC criterion is widely accepted and has been shown to perform well in mixture problems (McLachlan and Peel, 2000; Schlattmann and Böhning, 1997). Ray and Lindsay (2008) suggest a selection criterion based upon a quadratic-risk approach and show that this criterion performs well in the mixture context. It is, however, remarkable that in their evaluations based upon simulation studies (Table 1, 2, and 4 in Ray and Lindsay, 2008) the conventional BIC criterion performs considerably well. A further refinement to reduce bias is to utilize the best BIC-selected model with a nonparametric bootstrap procedure, resulting in a bootstrap adjusted estimator which has performed well in examples and in limited simulation studies.

On the negative side it should be noted that the suggested approach is computationally intensive and needs computational skill in computing the mixture model maximum likelihood estimator correctly. The gain in reducing the bias needs to be seen against this enormous computational burden. In addition, standard errors are occasionally large with the mixture model approach whereas Chao's estimator retains a rather low standard error. It will be left to future research to compare this mixture approach to simpler nonparametric procedures such as generalizations of Zelterman's or Chao's estimator. In addition, it might be valuable to include the penalized nonparametric maximum likelihoodapproach suggested by Wang and Lindsay (2005) in such comparisons.

Appendix

Mixture Maximum Likelihood Theory

The benefit of working with a mixture model of zero-truncated Poisson densities $f_+(j,\lambda_\ell) = Po(j;\lambda_\ell)/[1-\exp(0;\lambda_\ell)]$

$$f_+(j;Q) = \sum_\ell q_\ell f_+(j,\lambda_\ell)]$$

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 $f_1, f_2, \ldots, f_m$ be their frequencies. Then, the log-likelihood with respect to $f_+(j;Q)$

$$\log L_+(Q) = \sum_j f_j \log\left(f_+(j;Q)\right)$$

is a *concave functional* on the set of *all* discrete probability distributions (though it is not concave on the set of all discrete probability measures with exactly k support points). This is the main reason for achieving the following global results. An important, analytical tool is the *gradient function* defined for any discrete distribution $Q = \begin{pmatrix} \lambda_1 & \ldots & \lambda_k \\ q_1 & \ldots & q_k \end{pmatrix}$ as

$$d(\lambda, Q) = \frac{1}{n}\sum_{j=1}^{m} f_j \frac{f_+(j,\lambda)}{f_+(j,Q)}$$

where $f_+(j,Q) = q_1 f_+(j,\lambda_1) + q_2 f_+(j,\lambda_2) + \ldots + q_k f_+(k,\lambda_k)$. With the help of the gradient function, the *nonparametric maximum likelihood estimator* (NPMLE) can be characterized. The general mixture maximum likelihood theorem (Lindsay, 1983, Böhning, 1982) states that for $\hat{Q} = \begin{pmatrix} \lambda_1 & \ldots & \lambda_k \\ q_1 & \ldots & q_k \end{pmatrix}$

$$\hat{Q} \text{ is NPMLE } \Leftrightarrow d(\lambda, \hat{Q}) \le 1 \quad \text{for all} \quad \lambda > 0. \tag{8}$$

In addition, $d(\lambda, \hat{Q}) = 1$ for $\lambda \in \{\hat{\lambda}_1, \ldots, \hat{\lambda}_k\}$, the set of all support points of $\hat{Q}$. 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).

The Boundary Problem

We can illustrate the usefulness of the mixture maximum likelihood theorem by showing the *boundary problem* (Wang and Lindsay, 2008; Mao and Lindsay, 2007). For the case of mixtures of zero-truncated Poisson densities we have

$$f_+(1;\hat{Q}) \ge f_1/n,$$

where $\hat{Q}$ is the NPMLE. To verify this result we note first that

$$\lim_{\lambda\to 0} \frac{\lambda^j/j!}{\exp(\lambda)-1} = \begin{cases} 1, & \text{if } j=1 \\ 0, & \text{otherwise} \end{cases}$$

so that

$$\lim_{\lambda\to 0} d(\lambda,\hat{Q}) = \lim_{\lambda\to 0} \frac{1}{n}\sum_{j=1}^{m} f_j \frac{f_+(j,\lambda)}{f_+(j,\hat{Q})} = \frac{f_1/n}{f_+(1,\hat{Q})}.$$

Since $d(\lambda,\hat{Q}) \le 1$ for all $\lambda > 0$ by (8) the claimed overestimation result $f_+(1;\hat{Q}) \ge f_1/n$ follows.

Algorithms

A variety of numerical algorithms exist to find the global maximum likelihood estimator, the *nonparametric maximum likelihood estimator* (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 maximumlikelihood solution conditional upon the number of mixture components k though there is no guarantee for a non-local solution. To proceed in the EM context we need the *complete data log-likelihood* which is given in this case as

$$\sum_{j=1}^{m} f_j \sum_{\ell=1}^{k} z_{j\ell} \log f_+(j, \lambda_\ell) + \sum_{j=1}^{m} f_j \sum_{\ell=1}^{k} z_{j\ell} \log q_\ell \tag{9}$$

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

$$e_{j\ell} = E(z_{j\ell} \mid f_j; q_\ell, \lambda_\ell, \ell = 1, \ldots k) = \frac{f_+(j, \lambda_\ell)\, q_\ell}{\sum_{i=1}^{k} f_+(j, \lambda_i)\, q_i}\,. \tag{10}$$

In the *M-step* new values $\hat{\lambda}_1, \ldots, \hat{\lambda}_k, \hat{q}_1, \ldots, \hat{q}_k$ are found which maximize the expected version of (9) leading to

$$\hat{q}_\ell = \frac{1}{n} \sum_{j=1}^{m} f_j e_{j\ell}, \quad \text{for} \quad \ell = 1, \ldots k \tag{11}$$

as new estimates for the weights. The new estimates $\hat{\lambda}_1, \ldots, \hat{\lambda}_k$ need to be found as solutions of

$$\hat{\lambda}_\ell = \frac{\sum_{j=1}^{\mathrm{v}} j n_j e_{j\ell}}{\sum_{j=1}^{\mathrm{v}} n_j e_{j\ell}} (1 - \mathrm{e}^{-\hat{\lambda}_\ell}), \quad \text{for} \quad \ell = 1, \ldots k\,. \tag{12}$$

Note that (12) does not provide a closed form solution for $\hat{\lambda}_\ell$, but rather suggests an iterative solution of the form $\hat{\lambda}_\ell^{\mathrm{new}} = \frac{\sum_{j=1}^{m} j n_j e_{j\ell}}{\sum_{j=1}^{m} n_j e_{j\ell}} (1 - \mathrm{e}^{-\hat{\lambda}_\ell^{\mathrm{old}}})$ which needs to be iterated until convergence. The EM algorithm can be largely improved upon if gradient function techniques are incorporated (Böhning, 2003).

Acknowledgements *The authors wish to thank the Editors of the Biometrical Journal for the opportunity to publish their research in this special topic issue. Special thanks also to an Associate Editor and two referees for their helpful comments to improve the paper.*

Conflict of Interests Statement
The authors have declared no conflict of interest.

References

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge.

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.* Chapman & Hall/CRC, Boca Raton.

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., 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.

Böhning, D. and Schön, D. (2005). Nonparametric maximum likelihood estimation of population size based on the counting distribution. *Journal of the Royal Statistical Society, Series C* **54**, 721–737.

Böhning, D. and Kuhnert, R. (2006). The Equivalence of Truncated Count Mixture Distributions and Mixtures of Truncated Count Distributions. *Biometrics* **62**, 1207–1215.

Böhning, D. (2008). A simple variance formula for population size estimators by conditioning. *Statistical Methodology* **5**, 410–423.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics* **24**, 2350–2383.

Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *Annals of Statistics* **30**, 927–961.

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

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43**, 783–791.

Chen, H., Chen, J., and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society: Series B* **63**, 19–29.

Del Rio Vilas, V. J., Sayers, R., Sivam, K., Pfeiffer, D. U., Guitian, J., and Wilesmith, J. W. (2005). A case study of capture-recapture methodology using scrapie surveillance data in Great Britain. *Preventive Veterinary Medicine* , 303–317.

Del Rio Vilas, V. J. and Böhning, D. (2008). Application of one-list capture-recapture models to scrapie surveillance data in Great Britain. *Preventive Veterinary Medicine* (in print).

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.

Harris, I. R. (1991). The estimated frequency of zero for a mixed Poisson distribution. *Statistics and Probability Letters* **12**, 371–372.

Hoinville, L. J., Hoek, A., Gravenor, M. B., and McLean, A. R. (2000). Descriptive epidemiology of scrapie in Great Britain: results of a postal survey. *Veterinary Record* **146**, 455–461.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, No. 260, 663–685.

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. Institute of Statistical Mathematics, Hayward.

Lindsay, B. G. and Roeder, K. (1987). A unified treatment of integer parameter models. *Journal of the American Statistical Association* **82**, 758–764.

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

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* **64**, Series A, 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.

Mao, C. X. and Lindsay, B. G. (2007). Estimating the number of classes. *Annals of Statistics* **35**, 917–930.

Mao, C. X. (2006). Inference on the number of species through geometric lower bounds. *Journal of American Statistical Association* **101**, 1663–1670.

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

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

McLean, A. R., Hoek, A., Hoinville, L. J., and Gravenor M. B. (1999). Scrapie transmission in Britain: a recipe for a mathematical model. *Proceedings of the Royal Society: Biological Sciences* **266**, No. 1437, 2531–2538.

Oremus, M. (2005). Personal communication.

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.

Ray, S. and Lindsay, B. G. (2008). Model selection in high dimensions: a quadratic-risk-based approach. *Journal of the Royal Statistical Society, Series B* **70**, 95–118.

Schlattmann, P. and Böhning, D. (1997). Contribution to a paper by Richardson and Green. *Journal of the Royal Statistical Society, Series B* **59**, 782–783.

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

Sivam, K., Baylis, M., Gravenor, M. B., Gubbins, S., and Wilesmith, J. W. (2003). Occurrence of scrapie in GB: results of a postal survey in 2002. *Veterinary Record*, 782–783.

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

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

Wang, J.-P. and Lindsay, B. G. (2005). A penalized nonparametric maximum likelihood approach to species richness estimation. *Journal of the American Statistical Association* **100**, 942–959.

Wang, J.-P. and Lindsay, B. G. (2008). An exponential partial prior for improving nonparametric maximum likelihood estimation in mixture models. *Statistical Methodology* **5**, 30–45.

Zelterman, D. (1988). Robust estimation in truncated discrete distributions with application to capture recapture experiments. *Journal of Statistical Planning and Inference* **18**, 225–237.

