Identifiability and inferential issues in capture-recapture experiments with heterogeneous detection probabilities

Alessio Farcomeni and Luca Tardella

November 12, 2012

Abstract

We focus on a capture-recapture model in which capture probabilities arise from an unspecified distribution F. We show that model parameters are identifiable based on the unconditional likelihood. This is not true with the conditional likelihood. We also clarify that consistency and asymptotic equivalence of maximum likelihood estimators based on conditional and unconditional likelihood do not hold. We show that estimates of the undetected fraction of population based on the unconditional likelihood converge to the so-called estimable sharpest lower bound and we derive a new asymptotic equivalence result. We finally provide theoretical and simulation arguments in favour of the use of the unconditional likelihood rather than the conditional likelihood especially when one is willing to infer on the sharpest lower bound.

Key Words: Binomial Mixture; Capture-Recapture; Identifiability; Conditional Likelihood; Complete Likelihood; Unconditional likelihood.

1 Introduction

In capture-recapture model M_h (Otis *et al.*, 1978) there are S capture occasions and the capture probability of each animal is constant over all occasions but is allowed to be different from animal to animal. The individual probability is assumed to be distributed according to some unknown F. Untestable parametric assumptions are usually needed in order to make inference about the unknown population size N. In this paper, we investigate the consequences of having no assumptions on F, and give some insights on what can be learned about N in this setting.

Our semiparametric framework has been considered in the traditional works of Burnham and Overton (1978) and Chao (1989), while likelihood-based approaches can be found in Norris and Pollock (1996) and Wang and Lindsay (2008). There are also Bayesian approaches in Basu and Ebrahimi (2001), Tardella (2002) and Farcomeni and Tardella (2010). However, none of the above can be considered a conclusive solution of the problem.

In recent papers different cautionary notes have been written to discuss troublesome effects of the presence of heterogeneity (Huggins and Yip, 2001; Hwang and Huggins, 2005) and warn against identifiability issues (Huggins, 2001; Link, 2003; Holzmann *et al.*, 2006; Mao, 2008). Surprisingly, we find out that some identifiability issues suffered by the most widely used approach based on the so-called conditional likelihood are not suffered by the unconditional likelihood when the appropriate model parameterization is used. The use of the unconditional likelihood under a suitable parameterization will be shown to lead to a classical notion of model identifiability

However, our positive identifiability result related to the uniqueness of the sampling distribution corresponding to each parameter configuration is not sufficient to guarantee consistency of the unconditional MLE for the population size parameter. Notice that in the capture recapture setting consistency is a peculiar concept, defined as an asymptotic property related to the behaviour of the estimators as N – one of the model parameters – grows to infinity.

In order to infer on the unknown fraction of unobserved units we will consider, as in Mao (2008), a uniquely defined lower bound of the probability that an individual is never captured, and, correspondingly, a lower bound on the true unknown population size, and show consistency to this *sharpest lower bound* under mild conditions.

Both the conditional and unconditional likelihood have been shown by Sanathanan (1972) to yield consistent estimators of the population size and to be asymptotically equivalent under suitable regularity conditions. We point out in the following that the usually invoked equivalence in Sanathanan (1972, Theorem 2, p. 147) is valid under a strong identifiability condition (Rao, 1965), [Theorem A2 in the original paper], which is not met in our setting for the conditional likelihood (Link, 2003). See also Tardella and Farcomeni (2009).

We show that both estimators (the one based on the conditional and the one based on the unconditional likelihood) yield consistent inference for the sharpest lower bound.

We finally argue through theoretical results and numerical evidence that for that lower bound parameter there is some useful extra information in the unconditional likelihood which cannot be gathered for finite N through the conditional likelihood.

The paper is organized as follows: in Section 2 we fix the notation and model setup and review the two different likelihood approaches as well as properties and relations between the corresponding MLE. In Section 3, after reviewing the nonidentifiability arguments of Link (2003) for the conditional likelihood, we prove identifiability of the unconditional sampling model under a suitable parameterization. In Section 4 we investigate more thoroughly the asymptotic unconditional likelihood behaviour showing how one can get consistency to the sharpest lower bound. In Section 5 we illustrate some inferential benefits of the unconditional likelihood approach over the conditional likelihood approach. In Section 6 we give some concluding remarks.

2 Model specification and alternative likelihoods

In a capture-recapture experiment one records the occurrence of trapping of N animals during a sequence of S trapping occasions. Each trapped animal is uniquely identified with a tag so that the whole capture history can be recorded. The goal is inferring the unknown population size N based on the capture histories of those animals which have been trapped at least once. Only a *random* part of the population will be actually observed. If we conventionally label the observed units as $1, 2, ..., n$ the remaining $N - n$ units correspond to animals for which S consecutive zeros should have been recorded.

Capture histories are collected in a binary $N \times S$ matrix $\boldsymbol{X}$, up to row index permutation. We denote with $\boldsymbol{x}_i = (x_{i1}, ..., x_{iS})$ $(i = 1, 2, ..., N)$ the sequence of binary capture history for the i-th animal.

Model M_h assumes independent and identically distributed (i.i.d) binary capture outcomes X_{ik} with subject-specific capture probability

$$p_i = \Pr(X_{ik} = 1 | p_i) \quad \forall k = 1, ..., S$$

and p_i is in turn assumed i.i.d. from an unknown distribution F with support in $[0, 1]$. Let $\boldsymbol{X}$ be an $N \times S$ binary matrix, with generic i-th row $\boldsymbol{X}_i = (X_{i1}, ..., X_{iS})$, with $X_{ij} = 1$ if the i-th subject has been captured at the j-th occasion. Let n_k denote the number of subjects which have been captured exactly k times, $n_0 = N - n$, $n = \sum_{k=1}^{S} n_k$, and

$$P_k(F) = \binom{S}{k} \int_0^1 p^k (1-p)^{(S-k)} F(dp) \qquad k = 0, 1, ..., S. \tag{1}$$

be the corresponding capture probability. Note that the quantity $n = \sum_{k=1}^{S} n_k$ is random.

We consider the sampling distribution of the count vector $\boldsymbol{n}_{obs} = (n_1, ..., n_k, ..., n_S)$, which is a sufficient statistic,

$$Pr(\boldsymbol{n}_{obs}; N, F) = \frac{N!}{\prod_{k=0}^{S} n_k!} \prod_{k=0}^{S} P_k(F)^{n_k} \tag{2}$$

depending on the unknown finite population size N and the distribution of unobservable capture probabilities F. The likelihood function expressed in terms of the unknown parameters is also referred to as *unconditional likelihood* to distinguish it from a related likelihood function called *conditional likelihood*, which assumes as sampling distribution the same vector of positive counts of units conditionally on the observed sample size n and on the fact that the observed animals are captured at least once:

$$Pr_{(c)}(\boldsymbol{n}_{obs}; F, n) = \frac{n!}{\prod_{k=1}^{S} n_k!} \frac{\prod_{k=1}^{S} P_k(F)^{n_k}}{(1 - P_0(F))^{\sum_{k=1}^{S} n_k}} \tag{3}$$

In this conditional perspective the number of observed units n is no longer a random quantity but it is assumed to be fixed. The likelihood functions corresponding to the two model specifications (2) and (3) are in fact closely related by the following factorization

$$L(N,F) = Pr\left(\boldsymbol{X}; N, F\right) \quad = \quad L_{(c)}(F) \times L_{(r)}(N,F). \tag{4}$$

The first factor $L_{(c)}(F)$ is basically a multinomial likelihood corresponding to S cell counts $(n_k, k = 1, 2, ..., S)$ and it is usually referred to as *conditional likelihood*

$$\begin{aligned} L_{(c)}(F) &= \frac{n!}{\prod_{k=1}^{S} n_k!} \frac{\prod_{k=1}^{S} P_k(F)^{n_k}}{(1-P_0(F))^{\sum_{k=1}^{S} n_k}} \\ &= \frac{n!}{\prod_{k=1}^{S} n_k!} \prod_{k=1}^{S} P_{(c),k}(F)^{n_k} \end{aligned} \tag{5}$$

It depends *only* on the F parameter through the conditional probabilities

$$P_{(c),k}(F) = \frac{P_k(F)}{1-P_0(F)} = \Pr\left\{ \boldsymbol{X}_i : \sum_{j=1}^{S} X_{ij} = k \middle| \sum_{j=1}^{S} X_{ij} > 0 \right\}, \quad k = 1, 2, ..., S$$

where, for any $k \geq 1$, $P_{(c),k}(F)$ denotes the conditional probability of capturing an animal exactly k times given that it will be captured at least once in the S occasions. The residual factor $L_{(r)}(N,F)$ corresponds to

$$\begin{aligned} L_{(r)}(N,F) &= \frac{N!}{(N-n)!n!} P_0(F)^{(N-n)} (1-P_0(F))^n \\ &= \frac{N!}{\left(N - \sum_{k=1}^{S} n_k\right)! \left(\sum_{k=1}^{S} n_k\right)!} P_0(F)^{N - \sum_{k=1}^{S} n_k} (1-P_0(F))^{\sum_{k=1}^{S} n_k} \end{aligned} \tag{6}$$

which can be also immediately recognized as a *binomial likelihood* and turns out to be a function of *both* N and F.

From the seminal work of Sanathanan (1972), the classical inferential approach consists in breaking down the estimation process into two consecutive steps:

- estimate F, or its corresponding parameters, through the conditional likelihood $L_{(c)}(F)$, obtaining

$$\hat{F}_c = \arg\max_{F} L_{(c)}(F).$$

- plug $\hat{F}_c$ in $P_0(F)$ and hence in the residual likelihood and obtain an estimate of the population size through $L_{(r)}(N, \hat{F}_c)$. In this step the estimation of the parameter of interest N is simplified since we are left with a standard binomial model structure with unknown size parameter for which maximizing the likelihood yields the explicit expression

$$\hat{N}_{P_0(\hat{F}_c)} = \arg\max_{N} L_{(r)}(N, \hat{F}_c) \cong \left[\frac{n}{1 - P_0(\hat{F}_c)} \right].$$

  i.e. the so-called Horvitz-Thompson estimator.

We define

$$N(P_0(\cdot)) \quad = \quad \frac{n}{1 - P_0(\cdot)} \tag{7}$$

which will be useful in the following since it allows to maximize the binomial residual likelihood for any fixed $P_0(\cdot)$.

2.1 The notion of identifiability

We believe that it is essential to clarify different notions of identifiability of model parameters. Since alternative inferential approaches and different likelihood functions can be used to draw inference we make explicit the following identifiability and non-identifiability concepts. We refer to Basu (1983) for the general definition of identifiability in statistical models and to Paulino and Pereira de Barganca (1994) for a wider review and references.

We say that a statistical model represented by the usual triple

$$\mathcal{M} = \{\mathcal{X}, f(x;\theta); \Theta\}$$

is identified when for any arbitrarily fixed true parameter θ^* generating the observed data $X \in \mathcal{X}$ at random according to $f(x;\theta^*)$ one has that there is no other $\theta' \in \Theta$ for which

$$f(x;\theta^*) = f(x;\theta') \qquad \forall x \in \mathcal{X}. \tag{8}$$

We stress the fact that the notion of identifiability is fundamentally related to the parameterization used to specify the family of distributions, rather than to the family itself. This explains why sometimes it is explicitly claimed that identification is "a property of the likelihood" (Kadane, 1975).

When there are two parameters say θ^* and θ' for which (8) holds they are termed observationally equivalent and in this case we say that the model parameter θ is not identified. Indeed when there are observationally equivalent parameters the likelihood function is necessarily flat over all equivalence classes defined as the subsets of Θ containing observationally equivalent parameters. The likelihood function $L(\theta) = f(x_{obs};\theta)$ cannot distinguish parameters within the same equivalence class.

In the original statistical model M_h the sampling distribution is specified as $f(n_{obs};(N,F)) = Pr(\boldsymbol{n}_{obs};N,F)$ and is defined for any vector $\boldsymbol{n}_{obs}$ in the set of non-negative integers $\mathcal{Z}_+^S$ and for any parameter $\theta = (N,F) \in \mathcal{N} \times \mathcal{P}$; the first component N lives in the positive integer set $\mathcal{N}$ and the other one in the space of all probability distributions in the unit interval. We can therefore formally specify model M_h as follows

$$\left\{\mathcal{Z}_+^S; f(n_{obs};(N,F)); \mathcal{N} \times \mathcal{M}_S\right\} \tag{9}$$

We will see that for the statistical model specified as in (9) there is some source of non-identifiability for the distributional part F of the parameter θ so that two distinct parameters (N,F_1) and (N,F_2) can be observationally equivalent provided that F_1 and F_2 have the same first S moments (see Section 3 for a detailed argument). Once acknowledged this source of non-identifiability a natural way to simplify the inferential problem is reformulating the model in terms of the first S moments of F, that is, reformulating the same family of distributions for the observable quantities as in (2) in terms of a simpler parameter space.

Indeed we will show in Theorem 1 that under the suitable moment-based parameterization the usual notion of identifiability holds.

On the other hand if we specify the sampling distribution conditionally on each unit being captured the resulting statistical model

$$\mathcal{M}_h^{(c)} = \left\{\mathcal{Z}_+^S; f_{(c)}(\boldsymbol{n}_{obs};F) = Pr_{(c)}(\boldsymbol{n}_{obs};F); \mathcal{P}\right\} \tag{10}$$

without any functional restriction on F suffers from a more severe problem of non-identifiability as neatly pointed out in Link (2003).

We finally stress that another inferential issue affects both models and is related to parameter estimability. The usual notion of estimability of a parameter is associated with that of consistency, i.e., the asymptotic behaviour of the inferential process with an arbitrarily growing number of observations. However in our *finite* population setting especially in the first specification the asymptotic behaviour of the estimators can be (and typically is) understood as N, a part of the unknown parameter, grows to ∞. Hence both the parameter and the observed frequencies jointly diverge and this makes our asymptotic investigation a little bit different from the standard setting.

2.2 Alternative model parameterization based on moments of F

We now introduce the alternative parameterization for the specification of M_h.

It has long been recognized in the binomial mixture literature (Rolph, 1968; Lindsay, 1995) and more recently in the capture-recapture context (Yoshida *et al.*, 1999; Tardella, 2002; Link, 2003) that the probabilities $\boldsymbol{P}(F)$ as in (1) are in one-to-one correspondence with the first S moments of F through the linear

relation

$$P_k(F) = \sum_{r=k}^{S} \frac{S!}{k!(r-k)!(S-r)!}(-1)^{(r-k)} m_r(F) \qquad k = 0, 1, ..., S \tag{11}$$

where $m_r(F) = \int_{[0,1]} p^r F(dp)$ is the ordinary r-th moment of F. Recall we denote with $\mathcal{P}$ the class of probability distributions with support in $[0, 1]$. The set $\mathcal{M}_S = \left\{ \boldsymbol{m} = (m_1(F), ..., m_S(F)) : m_r(F) = \int_0^1 p^r F(dp),\ r = 1, ..., S, F \in \mathcal{P} \right\}$ is called the S-truncated moment space.

Hence one can represent more parsimoniously the model structure in terms of moment parametrization as follows

$$\mathcal{M}_h = \{ Pr(\boldsymbol{n}_{obs}; N, \boldsymbol{m}); \boldsymbol{m} \in \mathcal{M}_S, N \in \mathcal{N} \} \tag{12}$$

The likelihood factorization (4) is rephrased as follows:

$$L(N, \boldsymbol{m}) \quad = \quad L_{(c)}(\boldsymbol{m}) \times L_{(r)}(N, \boldsymbol{m}) \tag{13}$$

where the moment vector parameter $\boldsymbol{m}$ lies in the S-dimensional convex body $\mathcal{M}_S$ with non empty interior. Analogously one can rephrase equivalently the conditional model in terms of moments

$$\mathcal{M}_h^{(c)} = \left\{ Pr_{(c)}(\boldsymbol{n}_{obs}; \boldsymbol{m}); \boldsymbol{m} \in \mathcal{M}_S \right\}.$$

In the following section we review the main argument leading to the non-identifiability of the conditional model specification and show that the same problem does not occur with the unconditional model reparameterized with the first S moments of F as in (12).

3 Identifiability issues

Link (2003) showed that the use of the conditional likelihood for estimating $P_0(F)$ suffers from non identifiability which eventually prevents from achieving valid inference for the unknown population size N.

For any mixing distribution F with corresponding $P_0(F) < 1$, there is at least an infinite collection of other mixing distributions $\{G_\gamma = \gamma F + (1-\gamma)\delta_0; \gamma \in (0, 1)\}$ for which $L_{(c)}(F) = L_{(c)}(G_\gamma)$, where δ_0 is the degenerate distribution at 0.

The first S moments of G_γ are linearly related to those of F by the relation $m_k(G_\gamma) = \gamma m_k(F)$ and hence $P_k(G_\gamma) = \gamma P_k(F)$ for any $k = 1, ..., S$ so that

$$1 - P_0(G_\gamma) = \sum_{k=1}^{S} P_k(G_\gamma) = \sum_{k=1}^{S} \gamma P_k(F) = \gamma(1 - P_0(F)).$$

The ratios defining the corresponding conditional probabilities are identical even if the first S moments differ:

$$P_{(c),k}(G_\gamma) = \frac{P_k(G_\gamma)}{1 - P_0(G_\gamma)} = \frac{P_k(F)}{1 - P_0(F)} = P_{(c),k}(F) \qquad k = 1, ..., S. \tag{14}$$

This means that there is structural non-uniqueness and potential unboundedness (degeneracy) of conditional maximum likelihood. In fact F and G_γ or, rather, m_F and m_{G_γ} will provide equally likely different conditional estimates $n/(1 - P_0(F))$ and $n/(1 - P_0(G_\gamma))$ so that with an arbitrarily small value of γ one can get an unbounded estimate of N.

Similar problems can be seen to occur when catchability is bounded below by a fixed constant, and even in other more restricted cases. Refer to Link (2003) for a more thorough discussion about non-identifiability of the conditional likelihood.

It is easy to argue that the same identifiability issue holds if one considers model $\mathcal{M}_h^{(c)}$ parameterized in terms of moments since the observational equivalence can be also derived directly from the proportional moments $m_k(G_\gamma) = \gamma m_k(F)$.

In the next subsection we show that the unconditional likelihood is not affected by the same kind of non-identifiability.

3.1 Full identifiability of the unconditional model specification

In this section we prove that using the unconditional model specification as in (12) overcomes the identifiability problem for a finite population size N. This amounts to say that one cannot find non trivial equivalence classes such as those found in (14).

Theorem 1. *Assume $N > 0$ and consider the parametric model*

$$\left\{\mathcal{Z}_+^S; f(n_{obs}; (N, F)); \mathcal{N} \times \mathcal{M}_S\right\}$$

In this model the parameter vector $(N, \boldsymbol{m})$ is fully identified.

Proof. It is easy to argue that the mapping from $(N, \boldsymbol{m})$ to the space of distributions over $\mathcal{Z}_+^S$ specified as in (2) is injective. In fact, if we take two different $(N, \boldsymbol{m}) \neq (N', \boldsymbol{m}')$ either $N \neq N'$ or $N = N'$ but $\boldsymbol{m} \neq \boldsymbol{m}'$. In the first case the distribution of $n = \sum_{k=1}^S n_k$ is different since it is supported over different sets of integers hence the joint distributions of n_{obs} cannot be identical. In the second case again the multinomial distributions of $(N-n, n_{obs})$ cannot be the same since the mapping $\boldsymbol{m} \to (P_0(\boldsymbol{m}), ..., P_S(\boldsymbol{m}))$ is linear hence continuous and invertible. □

In order to get a more direct feeling of the different behaviour of the two likelihoods in Tardella and Farcomeni (2009) it is shown how one can perform simple numerical checks and see that the unconditional likelihood $L(N, G_\gamma)$ does not appear flat with $\gamma \in (0, 1)$ as is the case with $L_{(c)}(G_\gamma)$ in Link (2003).

Unfortunately, the positive result in Theorem 1 does not overcome all the inferential difficulties. As a matter of fact the likelihood surface can appear near flat on the log-scale on parameter regions derived from the conditional probability equivalence classes.

Indeed in a personal communication W. A. Link has pointed out that if one restrict the attention to the distributions belonging to conditional probability equivalence classes the conditional likelihood is constant and all that is left to help distinguishing among them is a single binomial with unknown index and succes rate. In fact that is another statistical model where inference is rather troublesome. Nonetheless there is still interest in comparing estimators, improving them and understanding their drawbacks (see DasGupta and Rubin (2005)). We can get insights considering the following example suggested by W. A. Link.

Example We consider the same example as in Link's response to Holzmann *et al.* (2006) with the following alternative parameter configurations:

- $N_1 = 384$, $F_1 \sim Beta(1/2, 3/2)$
- $N_2 = 256$, F_2 discrete with two point masses at $(1/4, 3/4)$ with probabilities $(3/4, 1/4)$.

For these distributions, taking $S = 4$ the corresponding first four moments differ being respectively $(1/4, 1/8, 5/64, 7/128)$ and $(3/8, 3/16, 15/128, 21/256)$. These are exactly proportional with ratio $\gamma = 3/2$ hence both conditional probabilities $P_j^{(c)}$ will coincide. On the other hand, the distribution of n will be different and precisely $Bin(384, 195/384)$ under (N_1, F_1) and $Bin(256, 195/256)$ under (N_2, F_2). Looking at the following figure one can appreciate that there is some difference in these two marginal distributions. We stress that the ratio of the conditional likelihoods evaluated the true moments of F_1 and F_2 will always be equal to the unity, while the ratio of the unconditional likelihoods will not. This is reflected by the fact that, even if the conditional probabilities coincide, the two models will generate rather different sample sizes n. In principle, then, one can find a threshold for the unconditional likelihood ratio to distinguish between the two models.

We performed a little simulation study to illustrate this point. We first generate data from a model with a parameter setting equal to (N_1, F_1) and decide whether the data have been generated by a model with parameter (N_1, F_1) or (N_2, F_2) using the maximized unconditional likelihood ratio using 1 as a threshold. In this first setting we select the true model 41% of the times. On the other hand when we simulate data from (N_2, F_2) we select the correct model in 76% of the cases. The actual percentages are obviously dependent on the threshold and the correspondence between the two types of errors determined by different values of the threshold can be read from the ROC curve in Figure 2. What is important here to note is that the conditional likelihood is not able to distinguish between these two cases, while the unconditional is: the use of the unconditional likelihood brings about discriminatory power which cannot be gained with the conditional likelihood. This does not anyway imply that one can safely test between the two models, and indeed our aim in this work is not to distinguish separate models.

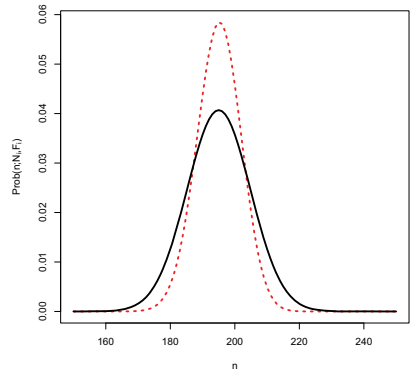

Figure 1: *Marginal distributions of the number of trapped units under parameter configuration* (N_1, F_1) *(solid line) and* (N_2, F_2) *(dotted line)*

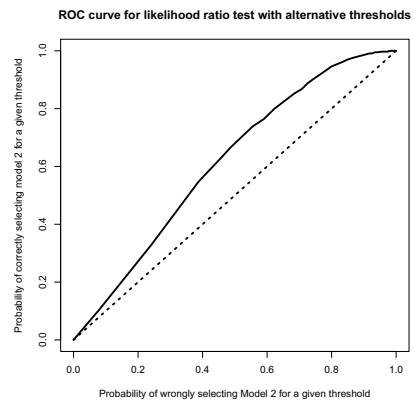

Figure 2: *Estimated ROC curve of the unconditional likelihood ratio test (solid line) to select between two alternative parameter configurations. The ROC curve is obtained by varying the likelihood ratio test threshold. 1000 simulated data under each parameter configurations has been used for the estimate. ROC curve of the conditional likelihood ratio test corresponds to the dotted segment joining the origin with the point (1,1).*

On the other hand the claimed identifiability of model parameters is somehow concealing another inferential issue which is related to the desired asymptotic behavior of MLE.

Interestingly, following from a personal communication by W. A. Link, we remark that our M_h model is a rare example of statistical parametric models where the sufficient statistics $\boldsymbol{n}_{obs}$ has a dimension which is smaller that the dimension of the fully identified parameter $\theta = (N, \boldsymbol{m})$.

4 Asymptotic inference and equivalence for the identified sharpest lower bound

Mao (2007, 2008) shows that despite the non-identifiability of the conditional likelihood corresponding to model $\mathcal{M}_h^{(c)}$ and the fact that one can always admit as MLE an arbitrarily large probability that an individual is never captured $P_0(F)$ it can be still of interest to draw inference on a non trivial, uniquely determined, sharpest lower bound of N, or equivalently of $P_0(F)$.

Given the true underlying (N_*, F_*), or, correspondingly in moment parameterization of (12), $(N_*, \boldsymbol{m}_*)$ where $\boldsymbol{m}_* = \boldsymbol{m}(F_*)$, the *sharpest lower bound* for $P_0(\cdot)$ is defined as

$$\phi_* = \phi_{F_*} = \phi_{\boldsymbol{m}_*} \quad = \quad \inf_{\boldsymbol{m} \in \mathcal{M}_S : \boldsymbol{P}_{(c)}(\boldsymbol{m}) = \boldsymbol{P}_{(c)}(\boldsymbol{m}_*)} P_0(\boldsymbol{m}). \tag{15}$$

In order to make inference on ϕ_{F_*} Mao (2008) shows how one can restrict the attention to a parametric family of identifiable distributions. He relies on the following identifiability notion related to a subfamily $\mathcal{A}$ of distributions

Definition 1. *A mixing distribution $G \in \mathcal{A}$ is strongly identifiable if the corresponding conditional probabilities $\boldsymbol{P}_{(c)}(G)$ cannot be reproduced by any other mixing distribution $H \in \mathcal{A}$, that is to say if $\boldsymbol{P}_{(c)}(G) = \boldsymbol{P}_{(c)}(H)$ for any $H \in \mathcal{A}$, then $G = H$.*

Mao (2008) shows that there exists a parametric family $\mathcal{K}$ of discrete mixing distributions with a suitable restriction on the number of positive masses which is strongly identifiable. Indeed each member of the $\mathcal{K}$ family is one-to-one with the space of all conditional probabilities $P_{(c),k}(F)$ spanned by the whole class $\mathcal{P}$ of probability distributions with support in $[0, 1]$ and attains the corresponding sharpest lower bound. In the rejoinder to Mao (2008), Holzmann and Munk (2008) show that if one assumes that the individual capture probability is larger than a fixed threshold, say $\delta > 0$, then the conditional likelihood approach consistently estimates $\phi_{\boldsymbol{m}_*}$ with the conditional MLE estimator

$$\hat{\phi}_{(c)} \quad = \quad \inf \left\{ P_0(\boldsymbol{m}') : \boldsymbol{m}' \in \arg \max_{\boldsymbol{m} \in \mathcal{M}_S} L_{(c)}(\boldsymbol{m}) \right\} \tag{16}$$

Note that $\hat{\phi}_{(c)}$ can be derived also as the conditional MLE restricted to the parametric identifiable sub-family $\mathcal{K}$ namely as the probability $P_0(\hat{G})$ where

$$\hat{G} \in \arg \max_{F \in \mathcal{K}} L_{(c)}(F)$$

We will show now that the unconditional estimator of $P_0(\boldsymbol{m}_*)$ is strictly related to the sharpest lower bound. In fact, it can be argued that

$$\hat{\phi} \quad = \quad P_0(\hat{\boldsymbol{m}}) \tag{17}$$

is an estimator of the sharpest lower bound.

Notice that $\hat{\boldsymbol{m}}$ corresponds to the unconditional likelihood maximization

$$L(\hat{N}, \hat{\boldsymbol{m}}) \geq L(N, \boldsymbol{m}) \qquad \forall N > 0 \quad \forall \boldsymbol{m} \in \mathcal{M}_S.$$

Under suitable conditions (slightly less general than the conditions in Holzmann and Munk (2008)) the sharpest lower bound ϕ_{F_*} can be consistently estimated by maximizing the unconditional likelihood. Precisely, if $(N_*, \boldsymbol{m}_*)$ is the true underlying parameter for model (12) and one considers the truncated moment sequence $\boldsymbol{m}_*^{LB}$ attaining the corresponding true sharpest lower bound

$$\phi_* = \phi_{\boldsymbol{m}_*} = \inf_{\boldsymbol{m} \in \mathcal{M}_S : P_{(c)}(\boldsymbol{m}) = P_{(c)}(\boldsymbol{m}_*)} P_0(\boldsymbol{m}) = P_0(\boldsymbol{m}_*^{LB}) \tag{18}$$

then $\boldsymbol{m}_*^{LB}$ can be consistently estimated.

We will see in Theorem 3 that both the conditional and unconditional likelihood estimators yield consistent estimators for the sharpest lower bound under suitable conditions. In this sense we extend somehow the equivalence result in Sanathanan (1972) beyond the regularity assumptions originally considered.

4.1 Likelihood asymptotics

Now we investigate formally the asymptotic behaviour of the unconditional likelihood evaluated at $(N(P_0(\hat{\boldsymbol{m}})), \hat{\boldsymbol{m}})$ in absence of the strong identifiability condition A.2 in Sanathanan (1972), which is clearly not met in the present case where no restriction is imposed on F.

The main result of the following theorem is in equation (20), which will be needed in the proof of (17) in the next section.

Recall that the binomial mixture probability space of all $P_k(\cdot)$ as in (1) and the truncated moment space are one-to-one, isomorphic and compact. In the following, to simplify the notation, we will use the same notation $\boldsymbol{P}(\boldsymbol{m})$ for the mapping from the truncated moment space. Moreover, we will consider that any sequence of estimators $\hat{\boldsymbol{m}}_{N_*}$ depends on an underlying true N_* and the corresponding estimates of the binomial mixture probabilities $\boldsymbol{P}(\hat{\boldsymbol{m}}_{N_*})$ always admit further subsequences converging to a limit, say $\tilde{\boldsymbol{m}}$. With a slight abuse of notation, we will denote (unless otherwise explicitly stated) the subsequence as $\hat{\boldsymbol{m}}_{N_*}$ and $\boldsymbol{P}(\hat{\boldsymbol{m}}_{N_*})$ and the corresponding limits $\tilde{\boldsymbol{m}}$ and $\boldsymbol{P}(\tilde{\boldsymbol{m}})$. Notice that $\tilde{\boldsymbol{m}}$ could be trivial (i.e. corresponding to a distribution δ_0 with all its mass at 0) and in that case the conditional probabilities are not well defined. Hence, with a slight abuse of notation again $\boldsymbol{P}_{(c)}(\tilde{\boldsymbol{m}})$ must be understood as an accumulation point for a suitable subsequence of well-defined conditional probabilities $\boldsymbol{P}_{(c)}(\hat{\boldsymbol{m}}_{N_*})$.

Theorem 2. *Assume the non trivial case where N_* and F_*, the parameters generating the observed counts $\boldsymbol{n}_{obs}$, are such that $N_* > 0$ and $\boldsymbol{m}_* = (m_1(F_*), \ldots, m_S(F_*))$ has its first component $m_1(F_*) > 0$. Let $\tilde{\boldsymbol{m}}$ denote an accumulation point of a subsequence of nontrivial moments $\hat{\boldsymbol{m}}_{N_*}$ for which the conditional probabilities $\boldsymbol{P}_{(c)}(\hat{\boldsymbol{m}}_{N_*})$ are well defined and converge to $\boldsymbol{P}_{(c)}(\tilde{\boldsymbol{m}})$. Then, denoting with KL the Kullback-Leibler divergence,*

$$\lim_{N_*\to\infty} \log \frac{L_{(r)}(N(P_0(\hat{\boldsymbol{m}}_{N_*})), \hat{\boldsymbol{m}}_{N_*})}{L_{(r)}(N_*, \boldsymbol{m}_*)} = -\frac{1}{2} \log \frac{P_0(\hat{\boldsymbol{m}}_{N_*})}{P_0(\tilde{\boldsymbol{m}})} \tag{19}$$

$$\lim_{N_*\to\infty} \frac{1}{N_*} \log \frac{L_{(c)}(\hat{\boldsymbol{m}}_{N_*})}{L_{(c)}(\boldsymbol{m}_*)} = -(1 - P_0(\boldsymbol{m}_*)) KL(\boldsymbol{P}_{(c)}(\tilde{\boldsymbol{m}}), \boldsymbol{P}_{(c)}(\boldsymbol{m}_*)) \tag{20}$$

$$\log \frac{L_{(c)}(\hat{\boldsymbol{m}}_{N_*})}{L_{(c)}(\boldsymbol{m}_*)} = O(N_*) \tag{21}$$

Furthermore if $P_0(\tilde{\boldsymbol{m}}) > 0$ then

$$\log \frac{L_{(r)}(N(P_0(\hat{\boldsymbol{m}}_{N_*})), \hat{\boldsymbol{m}}_{N_*})}{L_{(r)}(N_*, \boldsymbol{m}_*)} = o(N_*^{-1}) \tag{22}$$

$$\log L(N(P_0(\hat{\boldsymbol{m}}_{N_*})), \hat{\boldsymbol{m}}_{N_*}) \approx \log L_{(c)}(\hat{\boldsymbol{m}}_{N_*}) \tag{23}$$

Proof. It is immediately argued that (22) and (21) are consequences of (19) and (20). Comparing (22) and (21) one also gets the asymptotic equivalence of the behaviour of the conditional and unconditional likelihood evaluated at the MLE as stated in (23).

First we prove (19). Consider $\log L_{(r)}(N_*, \boldsymbol{m}_*)$. Indeed

$$\begin{aligned} \log L_{(r)}(N_*, \boldsymbol{m}_*) &= \log N_*! - \log(N_* - n)! - \log n! \\ &\quad + (N_* - n) \log P_0(\boldsymbol{m}_*) + n \log(1 - P_0(\boldsymbol{m}_*)) \\ &= (\star) - \log n! + (N_* - n) \log P_0(\boldsymbol{m}_*) + n \log(1 - P_0(\boldsymbol{m}_*)) \end{aligned}$$

where for the sum of the first two log-factorial terms, denoted with $(\star)$ we make use of the following Stirling's approximation

$$\log x! \cong x \log x - x + \frac{1}{2} \log x + o(\log(x)),$$

which is better and better as x grows, suppressing the $o(\log(x))$ summand in the following; hence we obtain

$$\begin{aligned}
(\star) &= N_* \log N_* - N_* + \frac{1}{2}\log N_* - (N_*-n)\log(N_*-n) + (N_*-n) - \frac{1}{2}\log(N_*-n) \\
&= N_* \log N_* - (N_*-n)\log(N_*-n) - n - \frac{1}{2}\log\frac{N_*-n}{N_*} \\
&= (N_*-n+n)\log N_* - (N_*-n)\log(N_*-n) - n - \frac{1}{2}\log\frac{N_*-n}{N_*} \\
&= (N_*-n)\log N_* - (N_*-n)\log(N_*-n) + n\log N_* - n - \frac{1}{2}\log\frac{N_*-n}{N_*} \\
&= -(N_*-n)\log\frac{N_*-n}{N_*} + n\log N_* - n - \frac{1}{2}\log\frac{N_*-n}{N_*}
\end{aligned}$$

Similarly, consider $\log L_{(r)}(N(P_0(\hat{\boldsymbol{m}}_{N_*})), \hat{\boldsymbol{m}}_{N_*})$. Indeed,

$$\begin{aligned}
\log L_{(r)}(N(P_0(\hat{\boldsymbol{m}}_{N_*})), \hat{\boldsymbol{m}}_{N_*}) &= \log N(P_0(\hat{\boldsymbol{m}}_{N_*}))! - \log(N(P_0(\hat{\boldsymbol{m}}_{N_*}))-n)! - \log n! \\
&\quad +(N_*(P_0(\hat{\boldsymbol{m}}_{N_*}))-n)\log P_0(\hat{\boldsymbol{m}}_{N_*}) + n\log(1-P_0(\hat{\boldsymbol{m}}_{N_*})) \\
&= (\star\star) - \log n! + (N_*(P_0(\hat{\boldsymbol{m}}_{N_*}))-n)\log P_0(\hat{\boldsymbol{m}}_{N_*}) \\
&\quad +n\log(1-P_0(\hat{\boldsymbol{m}}_{N_*}))
\end{aligned}$$

where we approximate and rearrange the factorial part $(\star\star)$, using (7), as follows

$$\begin{aligned}
(\star\star) &= N(P_0(\hat{\boldsymbol{m}}_{N_*}))\log N(P_0(\hat{\boldsymbol{m}}_{N_*})) - N(P_0(\hat{\boldsymbol{m}}_{N_*})) + \frac{1}{2}\log N(P_0(\hat{\boldsymbol{m}}_{N_*})) \\
&\quad -(N(P_0(\hat{\boldsymbol{m}}_{N_*}))-n)\log(N(P_0(\hat{\boldsymbol{m}}_{N_*}))-n) + (N(P_0(\hat{\boldsymbol{m}}_{N_*}))-n) - \frac{1}{2}\log(N(P_0(\hat{\boldsymbol{m}}_{N_*}))-n) \\
&= N(P_0(\hat{\boldsymbol{m}}_{N_*}))\log N(P_0(\hat{\boldsymbol{m}}_{N_*})) - (N(P_0(\hat{\boldsymbol{m}}_{N_*}))-n)\log(N(P_0(\hat{\boldsymbol{m}}_{N_*}))-n) - n \\
&\quad -\frac{1}{2}\log\frac{N(P_0(\hat{\boldsymbol{m}}_{N_*}))-n}{N(P_0(\hat{\boldsymbol{m}}_{N_*}))} \\
&= (N(P_0(\hat{\boldsymbol{m}}_{N_*}))-n+n)\log N(P_0(\hat{\boldsymbol{m}}_{N_*})) - (N(P_0(\hat{\boldsymbol{m}}_{N_*}))-n)\log(N(P_0(\hat{\boldsymbol{m}}_{N_*}))-n) - n \\
&\quad -\frac{1}{2}\log\frac{N(P_0(\hat{\boldsymbol{m}}_{N_*}))-n}{N(P_0(\hat{\boldsymbol{m}}_{N_*}))} \\
&= (N(P_0(\hat{\boldsymbol{m}}_{N_*}))-n)\log N(P_0(\hat{\boldsymbol{m}}_{N_*})) - (N(P_0(\hat{\boldsymbol{m}}_{N_*}))-n)\log(N(P_0(\hat{\boldsymbol{m}}_{N_*}))-n) + \\
&\quad +n\log N(P_0(\hat{\boldsymbol{m}}_{N_*})) - n - \frac{1}{2}\log\frac{N(P_0(\hat{\boldsymbol{m}}_{N_*}))-n}{N(P_0(\hat{\boldsymbol{m}}_{N_*}))} \\
&= -(N(P_0(\hat{\boldsymbol{m}}_{N_*}))-n)\log\frac{N(P_0(\hat{\boldsymbol{m}}_{N_*}))-n}{N(P_0(\hat{\boldsymbol{m}}_{N_*}))} + n\log N(P_0(\hat{\boldsymbol{m}}_{N_*})) + \\
&\quad -n - \frac{1}{2}\log\frac{N(P_0(\hat{\boldsymbol{m}}_{N_*}))-n}{N(P_0(\hat{\boldsymbol{m}}_{N_*}))} \\
&= -\frac{nP_0(\hat{\boldsymbol{m}}_{N_*})}{1-P_0(\hat{\boldsymbol{m}}_{N_*})}\log P_0(\hat{\boldsymbol{m}}_{N_*}) + n\log\frac{n}{1-P_0(\hat{\boldsymbol{m}}_{N_*})} - n - \frac{1}{2}\log P_0(\hat{\boldsymbol{m}}_{N_*}) \\
&= -n\frac{P_0(\hat{\boldsymbol{m}}_{N_*})}{1-P_0(\hat{\boldsymbol{m}}_{N_*})}\log P_0(\hat{\boldsymbol{m}}_{N_*}) + n\log n - n\log(1-P_0(\hat{\boldsymbol{m}}_{N_*})) - n - \frac{1}{2}\log P_0(\hat{\boldsymbol{m}}_{N_*})
\end{aligned}$$

Now the residual log-likelihood ratio in (19) is

$$\begin{aligned}
\log\frac{L_{(r)}(N(P_0(\hat{\boldsymbol{m}}_{N_*})), \hat{\boldsymbol{m}}_{N_*})}{L_{(r)}(N_*, \boldsymbol{m}_*)} &= [(\star\star)-(\star))] + (N(P_0(\hat{\boldsymbol{m}}_{N_*}))-n)\log P_0(\hat{\boldsymbol{m}}_{N_*}) + \\
&\quad n\log(1-P_0(\hat{\boldsymbol{m}}_{N_*})) - (N_*-n)\log P_0(\boldsymbol{m}_*) - n\log(1-P_0(\boldsymbol{m}_*)) \\
&= [(\star\star)-(\star))] + n\frac{P_0(\hat{\boldsymbol{m}}_{N_*})}{1-P_0(\hat{\boldsymbol{m}}_{N_*})}\log P_0(\hat{\boldsymbol{m}}_{N_*}) \\
&\quad +n\log(1-P_0(\hat{\boldsymbol{m}}_{N_*})) - (N_*-n)\log P_0(\boldsymbol{m}_*) - n\log(1-P_0(\boldsymbol{m}_*))
\end{aligned}$$

which simplifies as follows considering $(\star)$ and $(\star\star)$

$$\begin{aligned}
& n\log\frac{n}{N_*} + (N_*-n)\log\frac{N_*-n}{N_*} - (N_*-n)\log P_0(\boldsymbol{m}_*) - n\log(1-P_0(\boldsymbol{m}_*)) \\
& -\frac{1}{2}\log\frac{P_0(\hat{\boldsymbol{m}}_{N_*})}{\frac{N_*-n}{N_*}} \\
= \;& n\log\frac{\frac{n}{N_*}}{1-P_0(\boldsymbol{m}_*)} + (N_*-n)\log\frac{\frac{N_*-n}{N_*}}{P_0(\boldsymbol{m}_*)} - \frac{1}{2}\log\frac{P_0(\hat{\boldsymbol{m}}_{N_*})}{\frac{N_*-n}{N_*}} \qquad (24)
\end{aligned}$$

Since as $N_* \to \infty$ we have $\frac{n}{N_*} \to 1-P_0(\boldsymbol{m}_*)$ and $\frac{N_*-n}{N_*} \to P_0(\boldsymbol{m}_*)$ hence we get that asymptotics of the first two summands can be obtained discussing the following possible subsequences

1. $\frac{n}{N_*} > 1 - P_0(\boldsymbol{m}_*)$ (and hence $\frac{N_*-n}{N_*} < P_0(\boldsymbol{m}_*)$)
2. $\frac{n}{N_*} < 1 - P_0(\boldsymbol{m}_*)$ (and hence $\frac{N_*-n}{N_*} < P_0(\boldsymbol{m}_*)$)
3. $\frac{n}{N_*} = 1 - P_0(\boldsymbol{m}_*)$ (and hence $\frac{N_*-n}{N_*} = P_0(\boldsymbol{m}_*)$).

In the first case the first summand in (24) converges to $+1$ while the second converges to -1. In the second case the first summand in (24) converges to -1 while the second converges to 1. The last case is trivial. Hence (19) is proved.

In order to prove (20) we start from the definition of the conditional likelihood

$$\begin{aligned}
\lim_{N_*\to\infty} \frac{1}{N_*} \log \frac{L_{(c)}(\hat{\boldsymbol{m}}_{N_*})}{L_{(c)}(\boldsymbol{m}_*)} &= \lim_{N_*\to\infty} \frac{1}{N_*} \sum_{k=1}^{S} n_k \log \frac{P_{(c),k}(\hat{\boldsymbol{m}}_{N_*})}{P_{(c),k}(\boldsymbol{m}_*)} \\
&= \lim_{N_*\to\infty} \sum_{k=1}^{S} \frac{n_k}{N_*} \log \frac{P_{(c),k}(\hat{\boldsymbol{m}}_{N_*})}{P_{(c),k}(\boldsymbol{m}_*)} \\
&= \sum_{k=1}^{S} P_k(\boldsymbol{m}_*) \log \frac{P_{(c),k}(\tilde{\boldsymbol{m}})}{P_{(c),k}(\boldsymbol{m}_*)} \\
&= (1 - P_0(\boldsymbol{m}_*)) \sum_{k=1}^{S} \frac{P_k(\boldsymbol{m}_*)}{1 - P_0(\boldsymbol{m}_*)} \log \frac{P_{(c),k}(\tilde{\boldsymbol{m}})}{P_{(c),k}(\boldsymbol{m}_*)} \\
&= (1 - P_0(\boldsymbol{m}_*)) \sum_{k=1}^{S} P_{(c),k}(\boldsymbol{m}_*) \log \frac{P_{(c),k}(\tilde{\boldsymbol{m}})}{P_{(c),k}(\boldsymbol{m}_*)} \\
&= -(1 - P_0(\boldsymbol{m}_*)) KL(P_{(c),k}(\tilde{\boldsymbol{m}}), P_{(c),k}(\boldsymbol{m}_*))
\end{aligned}$$

Notice that $m_1(F_*) > 0$ necessarily yields that $P_k(\boldsymbol{m}_*) > 0$. □

4.2 Consistent estimator of the sharpest lower bound

We now prove the consistency of the unconditional MLE estimator of the sharpest lower bound under suitable compactness restrictions. In Section 5 we argue that when N_* is finite the unconditional likelihood can be a much better choice for deriving inference even if there is asymptotic equivalence.

Theorem 3. *Suppose that the true population size is N_* and the truncated moment sequence $\boldsymbol{m}_* = (m_1(F_*), \ldots, m_S(F_*))$ corresponds to the true capture probability distribution F_*. We assume that the support of any F is restricted to $[L, U] \subset [0, 1]$ with $0 < L < U < 1$. If $(\hat{N}, \hat{\boldsymbol{m}}_{N_*}) \in \arg\max_{N,m} L(N, \boldsymbol{m})$ and we consider any sequence $\hat{\boldsymbol{m}}_{N_*} = \hat{\boldsymbol{m}}[N_*]$ of maximizers when $N_* \to \infty$ then we get that*

$$P_0(\hat{\boldsymbol{m}}_{N_*}) \to \phi_{\boldsymbol{m}_*} \tag{25}$$

$$\hat{\boldsymbol{m}}_{N_*} \to \boldsymbol{m}^{LB} \tag{26}$$

where $\boldsymbol{m}^{LB}$ denotes the truncated moment sequence achieving the sharpest lower bound as in (15)

Proof. Since the truncated moment space is compact, the set $\arg\max_{N,m} L(N, \boldsymbol{m})$ is always non empty. The assumptions on the support of F yield estimates $(\hat{N}, \hat{\boldsymbol{m}}_{N_*}) \in \arg\max_{N,m} L(N, \boldsymbol{m})$ such that $0 < (1-U)^S \leq P_0(\hat{\boldsymbol{m}}_{N_*}) \leq (1-L)^S < 1$. Furthermore, under the above assumptions, the space of conditional probabilities $\boldsymbol{P}_{(c)}(m)$ becomes compact hence also the conditional MLE always yields estimates bounded in the same interval.

We start proving that, if $\tilde{\boldsymbol{m}}$ is an accumulation point of the sequence $\hat{\boldsymbol{m}}_{N_*}$ then it must be

$$KL(\boldsymbol{P}_{(c)}(\tilde{\boldsymbol{m}}), \boldsymbol{P}_{(c)}(\boldsymbol{m}_*)) = 0 \tag{27}$$

First, note that, if $(N(P_0(\hat{\boldsymbol{m}})), \hat{\boldsymbol{m}})$ maximizes the unconditional likelihood and $N(P_0(\hat{\boldsymbol{m}}))$ maximizes the residual likelihood $L_{(r)}(\cdot, \hat{\boldsymbol{m}})$ when $\hat{\boldsymbol{m}}$ is held fixed, we have

$$L(N(P_0(\hat{\boldsymbol{m}})), \hat{\boldsymbol{m}}) \geq L(N(P_0(\boldsymbol{m})), \boldsymbol{m}) \qquad \forall m \tag{28}$$

From (19), (20), (28) and the fact that Kullback-Leibler divergence is non negative one can obtain the following chain of inequalities

$$\begin{aligned} 0 &\leq \lim_{N_*\to\infty} \frac{1}{N_*} \log \frac{L_{(r)}(N(P_0(\hat{\boldsymbol{m}}_{N_*})), \hat{\boldsymbol{m}}_{N_*})}{L_{(r)}(N_*, \boldsymbol{m}_*)} + \frac{1}{N_*} \log \frac{L_{(c)}(\hat{\boldsymbol{m}}_{N_*})}{L_{(c)}(\boldsymbol{m}_*)} \\ &= -KL(\boldsymbol{P}_{(c)}(\tilde{\boldsymbol{m}}), \boldsymbol{P}_{(c)}(\boldsymbol{m}_*)) \\ &\leq 0. \end{aligned}$$

which proves (27). We have used the fact that $P_0(\tilde{\boldsymbol{m}}) \geq (1-U)^S > 0$ to ensure the appropriate asymptotic negligibility of the residual likelihood part resulting from (19)

From (27) it follows that any converging subsequence of any subsequence of the conditional probabilities must converge to the same limit $\boldsymbol{P}_{(c)}(\boldsymbol{m}_*)$ hence, for the whole original sequence $\boldsymbol{P}_{(c)}(\hat{\boldsymbol{m}}_{N_*})$ it must be

$$\boldsymbol{P}_{(c)}(\tilde{\boldsymbol{m}}) = \lim_{N_*\to\infty} \boldsymbol{P}_{(c)}(\hat{\boldsymbol{m}}_{N_*}) = \boldsymbol{P}_{(c)}(\boldsymbol{m}_*). \tag{29}$$

Let us consider the conditional likelihood $L_{(c)}(\boldsymbol{m})$ and the corresponding conditional likelihood estimator $\hat{\boldsymbol{m}}^{(c)}$ such that $L_{(c)}(\hat{\boldsymbol{m}}^{(c)}) \geq L_{(c)}(\boldsymbol{m})$ for all $\boldsymbol{m} \in \mathcal{M}_S$. Since the corresponding sharpest lower bounds $\phi(\hat{\boldsymbol{m}}^{(c)})$ are consistent estimators of the true sharpest lower bound under the assumptions of the theorem (Holzmann and Munk, 2008), there exists a sequence of conditional likelihood estimates such that $\phi(\hat{\boldsymbol{m}}^{(c)}) \to \phi_{\boldsymbol{m}_*}$. From the well known monotonicity relation (see also (32)) we get $P_0(\hat{\boldsymbol{m}}_{N_*}) \leq P_0(\hat{\boldsymbol{m}}^{(c)}) \leq \phi(\hat{\boldsymbol{m}}^{(c)}) \to \phi_{\boldsymbol{m}_*}$ and hence

$$P_0(\tilde{\boldsymbol{m}}) \leq \phi_{\boldsymbol{m}_*}. \tag{30}$$

On the other hand

$$P_0(\tilde{\boldsymbol{m}}) \geq \phi_{\boldsymbol{m}_*} \tag{31}$$

by definition of $\phi_{\boldsymbol{m}_*}$ and (29).

By (30) and (31) we get consistency as in (25). Finally, combining (25) with (29) and (18) for any accumulation point of suitable subsequences of the unconditional MLE it must be

$$\boldsymbol{P}(\tilde{\boldsymbol{m}}) = \boldsymbol{P}(\boldsymbol{m}^{LB}).$$

From the one-to-one correspondence between the truncated moment space and the space of mixture of binomial probabilities (26) is proved. □

Holzmann and Munk (2008) assume only a lower bound $L > 0$ the support of F while here we we also need to restrict the upper bound $U < 1$. The consistency result can be generalized under a milder moment condition such that $0 < L < m_1 < U < 1$, which is weaker than the lower bound on p. Details are omitted for reasons of space.

We now come back to Example 1 to show some other arguments which can help understanding the role of the residual likelihood and the origin of the sharpest lower bound inference

Example The following issue has been raised by W. A. Link in a personal communication where the discussion was focussed on the inferential limit of the likelihood function for model M_h. Quoting from his note "inference will rarely be limited to a choice of A) $N_1 = 384, F_1 \sim B(1/2, 3/2)$ versus B) $N_2 = 256, F_2$ is discrete with two point masses at $(1/4, 3/4)$ with probabilities $(3/4, 1/4)$. Instead, suppose that we know that $F = F_1$ or F_2, but N is unknown. Either way, $n|N \sim B(N, p)$ but with $p = p_\beta = 195/384$ if $F = F_1$, and with $p = p_2 = 195/256$ if $F = F_2$. Letting X denote an indicator for $F = F_1$, the full likelihood is then proportional to

$$L(X, N) = \binom{N}{n} \{p_\beta^n (1-p_\beta)^{N-n}\}^X \{p_2^n (1-p_2)^{N-n}\}^{1-X}$$

it can be shown that this is maximized by $X = 0$ and $N \approx n/p2$, for all n. Thus, regardless of the data, the 2-point mixture model is favored. And why? for no other reason than that $p_2 > p$." This seemingly unsatisfactory likelihood behaviour can reveal from a different perspective the motivation behind the use of the sharpest lower bound. In fact, if we assume no restriction on the success probability and we follow the likelihood principle we must agree that the best inference that we can do with a single number of successes $n = 195$ from an unknown number of trials with completely unknown probability (not limited to only two

values) is to guess that they have been most likely yielded by $\hat{N} = 195$ and success probability equals to 1. Though partially disappointing there is nothing wrong with that. When we have the choice of the success probability restricted within a single equivalence class of distributions for which the conditional probabilities are the same, but they correspond to different binomial success probabilities (actually the probabilities p_2 and p_β correspond to the probability of never being observed) this can help us understanding once again why the unconditional likelihood is geared towards the sharpest lower bound.

5 Advantages of the unconditional likelihood in finite samples

5.1 Theoretical arguments in favor of the unconditional likelihood

It is argued in Sanathanan (1972), and very easy to show, that

$$P_0(\hat{\boldsymbol{m}}) \leq P_0(\hat{\boldsymbol{m}}^{(c)}) \tag{32}$$

for any $\hat{\boldsymbol{m}}$ maximizing the unconditional and any $\hat{\boldsymbol{m}}^{(c)}$ maximizing the conditional likelihood.

Since (32) holds for any $\hat{\boldsymbol{m}}^{(c)}$ maximizing the conditional likelihood, it turns out that

$$\hat{\phi} = P_0(\hat{\boldsymbol{m}}) \leq \inf_{\boldsymbol{m} \in \mathcal{M}_S : P_{(c)}(\boldsymbol{m}) = P_{(c)}(\hat{\boldsymbol{m}}^{(c)})} P_0(\hat{\boldsymbol{m}}^{(c)}) \leq \hat{\phi}_{(c)}. \tag{33}$$

This fact reveals that the unconditional likelihood always provides an estimate which represents an approximation from below of the sharpest lower bound. Indeed we will see that for finite N_* estimating ϕ_0 from below yields a more stable behaviour than the corresponding conditional MLE.

Another relevant argument against the use of the conditional maximum likelihood estimator of $\phi_{\boldsymbol{m}_*}$ is that it can (and does sometimes) occur at the boundary, i.e., $P_0(\hat{\boldsymbol{m}}^{(c)})$ can be arbitrarily close to 1, leading to an unbounded estimate for the population size N. This has been already formally argued and empirically experienced also in parametric subclasses (Mao and You, 2009).

The fundamental reason is of topological nature: the mapping from the truncated moment space to the mixture of binomial probabilities is continuous and maps a compact subset into another compact subset, while the mapping from the truncated moment space to the conditional probabilities corresponding to the mixture of binomial probability space is not continuous and maps to an open set. Discontinuities are in fact encountered for those sequences of truncated moments corresponding to sequences for which the first moment vanishes. In such cases the limit points of the conditional probabilities can be indeterminate. This can explain why maximizing the conditional likelihood can in practice yield an arbitrarily large estimate of $P_0(F)$.

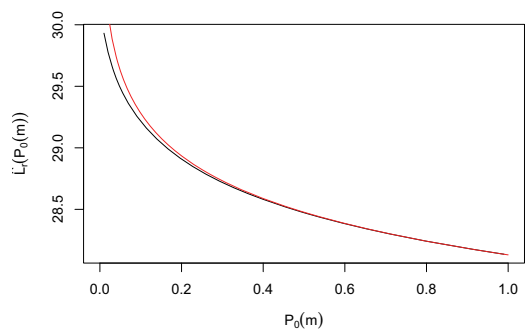

Figure 3: *Residual maximized likelihood* $L_{(r)}(N(P_0(\boldsymbol{m})), \boldsymbol{m})$ *when* $n = 10$

Remark One might wonder whether the unconditional MLE, unlike the conditional MLE, enjoys a theoretical non-degeneracy property. Although one verifies in practice that maximizing the unconditional likelihood yields only very rarely unstable/degenerate estimates $1 - P_0(\hat{\boldsymbol{m}})$ and hence an unbounded estimate of the unknown population size $\hat{N}$ one has to acknowledge that the penalization due to the presence of the residual likelihood cannot prevent that to happen. In fact it is easy to check either numerically (see Figure 3) or with formal arguments similar to those used in the proof of Theorem 1 that the residual likelihood $L_{(r)}(N(P_0(\boldsymbol{m})), \boldsymbol{m})$ evaluated at the maximizing $N(P_0(\boldsymbol{m})$ represents a penalization term which does not vanish as $P_0(\boldsymbol{m}) \to 1$. In fact it is bounded from below by a constant which is achieved with a functional form proportional to $P_0(\boldsymbol{m})^{-\frac{1}{2}}$ as $P_0(\boldsymbol{m}) \to 1$.

5.2 Unconditional likelihood performance in estimating the sharpest lower bound

In this subsection we consider the estimable sharpest lower bound as the quantity of interest and empirically show that the performance of the two likelihood approaches can differ substantially when N_* is finite. We do not compare the conditional and unconditional likelihood in terms of the corresponding estimates of N, as we have clarified that $P_0(\boldsymbol{m}_*)$ and hence N cannot be consistently estimated. One can consider the unconditional estimate $\hat{N}$ or the corresponding $\frac{n}{1-P_0(\hat{\boldsymbol{m}})}$ a *partial* inference on tre true N. Indeed given that both likelihoods yield consistent estimates for the sharpest lower bound we will focus on comparing the estimators of the sharpest lower bound. It is crucial to verify which one (if any) gives a better performance in terms of the ability to recover the underlying ϕ_*. We will see that the unconditional likelihood approach can improve the quality of inference in terms of a reduced frequency of boundary estimates, reduced bias (typically positive), and a much smaller mean square error.

We consider the same benchmark simulation plan used in Pledger (2005), also revisited by other authors (e.g., Mao and You (2009); Dorazio and Royle (2005)). The simulation settings considered are divided in three groups (A,B,C), and there are six settings for each group, numbered from 1 to 6. Overall there are 18 different underlying distributions F ranging from discrete distributions to continuous distributions from different parametric families. We refer the reader to Pledger (2005) for a detailed description of those distributions. From each distribution we generated $B = 100$ simulated data sets using as a true population size $N_* = 100$ and $N_* = 1000$. Hence the simulation plan considered $18 \times 2 \times 100$ simulated data and the corresponding estimates.

In order to implement the conditional and unconditional MLE we have reparameterized the model (12) in terms of canonical moments (Dette and Studden, 1997), an unconstrained one-to-one transformation of the moments. For a similar transformation see for instance Tardella (2002) and Farcomeni and Tardella (2010). This allowed us to use standard maximization routines in an unconstrained space. For the unconditional likelihood we have avoided the maximization over the integer N by using a reparameterized version of the profile likelihood $L(N(P_0(\boldsymbol{m})), \boldsymbol{m})$, where $N(P_0(\boldsymbol{m}))$ is defined as in (7).

We estimate the root mean square error (RMSE) of each estimator with $\sqrt{1/B \sum_{b=1}^{B} (\hat{\phi}_b - \phi_{\boldsymbol{m}_*})^2}$, where $\hat{\phi}_b$ represents the estimator obtained with the b-th simulated dataset (either the conditional or unconditional likelihood estimator) and $B = 100$ is the total number of replications. As relative efficiency measure of unconditional MLE versus conditional MLE we report the ratio between the two RMSE in estimating the sharpest lower bound. We also report for each simulation setting the odds corresponding to the sharpest lower bound $\phi_{\boldsymbol{m}_*}/(1-\phi_{\boldsymbol{m}_*})$ and the ratio of RMSE for estimating the odds. In the ratios, we put the RMSE corresponding to the unconditional likelihood at numerator ($RMSE_{\hat{\phi}}$) and the one corresponding to the conditional likelihood at denominator. Therefore, a ratio greater than 1 indicates a better performance of the unconditional likelihood estimator. Finally, for both estimators we report the estimated probability of obtaining an estimate larger than 0.95. Results are displayed in Table 1 when the true population size is set as $N_* = 100$ and in Table 2 when $N_* = 1000$. We denote with $\phi_{\boldsymbol{m}_*}$ and τ_* the true sharpest lower bound and true odds corresponding to each simulation setting.

When $N_* = 100$ and the conditional likelihood approach is used we have experienced a basically unbounded estimate of of the sharpest lower bound in a fraction of the simulated dataset ranging from 6% to 38%. With the unconditional likelihood this fraction is almost always null, and always drastically reduced. The bias of the estimator is typically positive (not shown). The unconditional likelihood estimator is substantially better than the conditional estimator both for the estimable lower bound in its original scale and in terms of the odds. It can be seen than the ratio between the estimated RMSE of the conditional and unconditional likelihood approach is always larger than 1, and often larger than 2.

When $N_* = 1000$ we still see a non negligible fraction of unbounded estimates with the conditional likelihood, while the unconditional likelihood never yields unbounded estimates. There advantage in terms

	$\phi_{\boldsymbol{m}_*}$	$\frac{RMSE_{\hat{\phi}}}{RMSE_{\hat{\phi}_{(c)}}}$	$\tau_* = \frac{\phi_{\boldsymbol{m}_*}}{1-\phi_{\boldsymbol{m}_*}}$	$\frac{RMSE_{\hat{\tau}}}{RMSE_{\hat{\tau}_{(c)}}}$	$Pr\{\hat{\phi}_{(c)} > 0.95\}$	$Pr\{\hat{\phi} > 0.95\}$
A1	0.40	2.08	0.67	4.57	0.30	0.02
A2	0.42	2.38	0.73	5.09	0.28	0.01
A3	0.46	1.77	0.84	3.35	0.28	0.02
A4	0.42	1.83	0.71	4.36	0.20	0.01
A5	0.44	1.84	0.79	> 250.00	0.28	0.02
A6	0.40	2.45	0.67	25.19	0.36	0.01
B1	0.24	2.87	0.31	220.95	0.26	0.00
B2	0.26	2.80	0.35	239.09	0.10	0.00
B3	0.33	2.10	0.49	3.17	0.35	0.04
B4	0.24	2.76	0.32	16.73	0.06	0.01
B5	0.32	2.38	0.47	176.53	0.38	0.00
B6	0.26	3.12	0.34	> 250.00	0.28	0.00
C1	0.34	1.81	0.52	2.89	0.26	0.05
C2	0.48	1.55	0.94	33.75	0.09	0.01
C3	0.35	2.02	0.54	3.15	0.33	0.04
C4	0.24	3.23	0.31	3.46	0.25	0.01
C5	0.33	2.43	0.49	> 250.00	0.29	0.01
C6	0.22	2.57	0.29	> 250.00	0.28	0.02

Table 1: *Performance of the unconditional and conditional MLE estimators under the 18 scenarios in Pledger (2005), when* $N_* = 100$. *The results are based on* $B = 100$ *replications.* $\hat{\phi}$ *is the estimator of the sharpest lower bound derived from the unconditional likelihood as in* (17). $\hat{\phi}_{(c)}$ *is the estimator of the sharpest lower bound derived from the conditional likelihood as in* (16).

of RMSE when using the unconditional likelihood is more or less of the same magnitude as when $N_* = 100$. We do not report further simulations for reasons of space, but we can mention that in our experiments the unconditional likelihood provides an advantage for $N_* < 10000$. For larger values of N_* the two approaches become essentially equivalent.

6 Concluding remarks

In this paper we have investigated from a theoretical perspective the controversial issue about the identifiability of the parameters involved in model M_h when no restrictive assumption is made on the distribution of the heterogeneous probabilities. We have shown how inference for model M_h through the unconditional likelihood can be carried out overcoming some of the identifiability concerns originally raised in Link (2003). Using the parametrization based on the first S moments of F and the corresponding unconditional likelihood the model as specified in (12) is fully identifiable. On the other hand we also pointed out that it is still not possible to obtain a consistent estimator of the true population size through the unconditional MLE. Indeed the unconditional MLE of the probability P_0 of never capturing an animal during all S trapping occasions is actually consistently estimating the sharpest lower bound rather than the fully identified probability of never capturing an individual. One could in principle consider to carry out inference within a Bayesian setting. However, the asymptotic equivalence of the MLE and Bayesian estimators would probably prevent consistent inference on P_0.

On the other hand we have shown that the classical unconditional MLE estimator of P_0 yields a consistent estimator for the sharpest lower bound under suitable regularity conditions. Similarly one can get a consistent estimator for the sharpest lower bound by computing the sharpest lower bound corresponding to the conditional MLE of the parameters as in (16). The conditional and unconditional likelihoods are asymptotically equivalent in this sense. This can explain the primary source of (possibly severe) negative bias when the estimate of the true population size N is at stake and one is trying to get an estimate of N_* with either one of the approaches. Nevertheless, suppose an upper bound for N or P_0 is roughly guessed, and (as typically the case in the ecological modelling) the upper bound is believed to be relatively small. In that case, inference based on the unconditional likelihood is technically sound since model parameters are fully identifiable and for this reason it could be preferred to inference based on the conditional likelihood. Other advantages have been also empirically illustrated with a benchmark simulation study where the unconditional likelihood is seen to provide more stable inference and a substantially smaller mean square error

	ϕ_*	$\frac{RMSE_{\hat{\phi}}}{RMSE_{\hat{\phi}_{(c)}}}$	$\tau_* = \frac{\phi_*}{1-\phi_*}$	$\frac{RMSE_{\hat{\tau}}}{RMSE_{\hat{\tau}_{(c)}}}$	$Pr\{\hat{\phi}_{(c)} > 0.95\}$	$Pr\{\hat{\phi} > 0.95\}$
A1	0.40	3.04	0.67	147.31	0.06	0.00
A2	0.42	2.64	0.73	31.46	0.04	0.00
A3	0.46	2.12	0.84	110.06	0.22	0.00
A4	0.42	2.38	0.71	31.92	0.02	0.00
A5	0.44	2.50	0.79	138.11	0.21	0.00
A6	0.40	3.01	0.67	76.58	0.13	0.00
B1	0.24	3.92	0.31	18.04	0.00	0.00
B2	0.26	4.72	0.35	11.45	0.00	0.00
B3	0.33	1.71	0.49	1.75	0.24	0.05
B4	0.24	4.61	0.32	9.95	0.00	0.00
B5	0.32	2.15	0.47	2.33	0.10	0.01
B6	0.26	2.89	0.34	23.77	0.02	0.00
C1	0.34	2.72	0.52	79.26	0.08	0.00
C2	0.48	3.05	0.94	22.18	0.03	0.00
C3	0.35	2.94	0.54	45.45	0.10	0.00
C4	0.24	3.67	0.31	50.31	0.01	0.00
C5	0.33	3.28	0.49	204.94	0.20	0.00
C6	0.22	3.43	0.29	44.60	0.03	0.00

Table 2: *Performance of the unconditional and conditional MLE estimators under the 18 scenarios in Pledger (2005), when* $N_* = 1000$. *The results are based on* $B = 100$ *replications.* $\hat{\phi}$ *is the estimator of the sharpest lower bound derived from the unconditional likelihood as in* (17). $\hat{\phi}_{(c)}$ *is the estimator of the sharpest lower bound derived from the conditional likelihood as in* (16).

when the sharpest lower bound is the parameter of interest.

Other inferential aspects in model M_h may deserve further investigation starting with the possibility of determining alternative estimators other than the unconditional MLE which can possibly consistently estimate the sharpest lower bound without any untestable compactness assumptions, i.e., without ruling out *a priori* small capture probabilities.

Although we acknowledge and agree on the importance of the consistency property for an estimator, the interesting point we highlight in this non-regular model is the following: there are examples of statistical models where identifiability and consistency – sometimes referred to as estimability – do not go together. Yet there is some partial information, possibly limited, which can be gained by inferring such identifiable models. Of course the reader must realize that as more and more units are observed inference will make us learn more and more on the sharpest lower bound $\phi(\boldsymbol{m}_*)$ rather than on the $P_0(\boldsymbol{m}_*)$ parameter.

Acknowledgements

The authors are deeply grateful to William A. Link for a stimulating and lively discussion.

References

A. P. Basu (1983). Icing the tails to limit theorems, lecture notes in economics and mathematical systems, 192. In: Samuel Kotz, Norman L. Johnson, and Campbell B. Read, eds., *Encyclopedia of statistical sciences. Vol. 4*, A Wiley-Interscience Publication, ix+657. John Wiley & Sons Inc., New York.

Sanjib Basu and Nader Ebrahimi (2001). Bayesian capture-recapture methods for error detection and estimation of population size: heterogeneity and dependence. *Biometrika*, **88, 1**, 269–279.

K. P. Burnham and W. S. Overton (1978). Estimation of the size of a closed population when capture probabilities vary among animals (corr: V68 p345). *Biometrika*, **65**, 625–634.

Anne Chao (1989). Estimating population size for sparse data in capture-recapture experiments. *Biometrics*, **45**, 427–438.

A. DasGupta and Herman Rubin (2005). Estimation of binomial parameters when both n, p are unknown. *Journal of Statistical Planning and Inference*, **130, 1–2**, 391–404.

Holger Dette and William J. Studden (1997). *The theory of canonical moments with applications in statistics, probability, and analysis*. John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.

Robert M. Dorazio and J. Andrew Royle (2005). Rejoinder to "the performance of mixture models in heterogeneous closed population capture-recapture". *Biometrics*, **61, 3**, 874–876.

A. Farcomeni and L. Tardella (2010). Reference Bayesian methods for alternative recapture models with heterogeneity. *TEST*, **19**, 187–208.

Hajo Holzmann and Axel Munk (2008). On the nonidentifiability of population sizes (rejoinder). *Biometrics*, **64, 3**, 977–979.

Hajo Holzmann, Axel Munk, and Walter Zucchini (2006). On identifiability in capture-recapture models. *Biometrics*, **62, 3**, 934–936.

R. Huggins (2001). A note on the difficulties associated with the analysis of capture-recapture experiments with heterogeneous capture probabilities. *Statistics & Probability Letters*, **54**, 147–152.

Richard M. Huggins and Paul S. F. Yip (2001). A note on nonparametric inference for capture-recapture experiments with heterogeneous capture probabilities. *Statistica Sinica*, **11, 3**, 843–853.

Wen-Han Hwang and Richard Huggins (2005). An examination of the effect of heterogeneity on the estimation of population size using capture-recapture data. *Biometrika*, **92, 1**, 229–233.

Joseph B. Kadane (1975). The role of identification in Bayesian theory. In: Stephen E. Fienberg and Arnold Zellner, eds., *Studies in Bayesian econometrics and statistics in honor of Leonard J. Savage*, 175–191. Elsevier/North-Holland [Elsevier Science Publishing Co., New York; North-Holland Publishing Co., Amsterdam].

Bruce G. Lindsay (1995). *Mixture Models: Theory, Geometry, and Applications*. Institute of Mathematical Statistics.

William A. Link (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics*, **59, 4**, 1123–1130.

C. X. Mao and N. You (2009). On comparison of mixture models for closed population capture-recapture studies. *Biometrics*, **65**, 547–553.

Chang Xuan Mao (2007). Estimating population sizes for capture-recapture sampling with binomial mixtures. *Comput. Stat. Data Anal.*, **51, 11**, 5211–5219.

Chang Xuan Mao (2008). On the nonidentifiability of population sizes. *Biometrics*, **64, 3**, 977–979.

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

D. L. Otis, K. P. Burnham, G. C. White, and D. R. Anderson (1978). *Statistical Inference From Capture Data on Closed Animal Populations*. Wildlife Monographs.

Carlos D. Paulino and Carlos A. Pereira de Barganca (1994). On identifiability of parametric statistical models. *Journal of the Italian Statistical Society*, **1, 3**, 125–151.

Shirley Pledger (2005). The performance of mixture models in heterogeneous closed population capture-recapture. *Biometrics*, **61, 3**, 868–876.

C. R. Rao (1965). *Linear Statistical Inference*. Wiley, New York.

John E. Rolph (1968). Bayes estimation of mixing distributions. *The Annals of Mathematical Statistics*, **39**, 1289–1302.

Lalitha Sanathanan (1972). Estimating the size of a multinomial population. *The Annals of Mathematical Statistics*, **43**, 142–152.

L. Tardella and A. Farcomeni (2009). Identifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Tech. Rep. 2*, Department of Statistics, Sapienza - University of Rome.

Luca Tardella (2002). A new Bayesian method for nonparametric capture-recapture models in presence of heterogeneity. *Biometrika*, **89, 4**, 807–817.

Ji-Ping Wang and Bruce G. Lindsay (2008). An exponential partial prior for improving nonparametric maximum likelihood estimation in mixture models. *Stat. Methodol.*, **5, 1**, 30–45.

O. Yoshida, J. G. Leite, and H. Bolfarine (1999). Stochastic monotonicity properties of Bayes estimation of the population size for capture-recapture data. *Statistics and Probability Letters*, **42**, 257–266.