On identifiability in capture-recapture models

Hajo Holzmann[1], Axel Munk[1] and Walter Zucchini[2]

Institute for Mathematical Stochastics[1] and Institute for Statistics and Econometrics[2]

Georg-August-University Göttingen, Germany

Abstract

We study the issue of identifiability of mixture models in the context of capture-recapture abundance estimation for closed populations. Such models are used to take account of individual heterogeneity in capture probabilities, but their validity was recently questioned by Link (2003) [*Biometrics* **59**, 1123–1130] on the basis of their non-identifiability. We give a general criterion for identifiability of the mixing distribution, and apply it to establish identifiability within families of mixing distributions that are commonly used in this context, including finite and beta mixtures. Our analysis covers binomial and geometrically distributed outcomes.

KEY WORDS: abundance estimation; capture-recapture; heterogeneity; identifiability; finite mixture; beta mixture.

1 Introduction

Capture-recapture methods are widely used in wildlife abundance estimation and also in fields such as epidemiology and quality control. They have been developed to estimate the size of both closed and open populations, but here, we restrict our attention to the former. For terminology and an overview of the methods see, e.g., Seber (1982).

An important issue in this context is the fact that, in many applications, the probability of capture/recapture differs among individuals in ways that are caused by factors that are not, or cannot be, observed (see, e.g., Borchers, Buckland and Zucchini, 2002, Section 11.3). Ignoring such heterogeneity can lead to substantial bias, and to inaccurate confidence intervals. One can address this problem by regarding the capture probabilities as realizations of a random variable, from which it follows that the number of animals captured in x out of T capture occasions, follows a mixture distribution (Burnham, 1972; Agresti, 1994; Norris and Pollock, 1995, 1996; Pledger, 2000, 2004; Dorazio and Royle, 2003). However, the use of mixture models raises the issue of identifiability (e.g. Huggins, 2001). Indeed Link (2003) concludes "Thus even with very large samples, the analyst will not be able to distinguish among reasonable models of heterogeneity, even though these yield quite distinct inferences about population size." Furthermore he gives examples to illustrate this statement, thereby casting doubt on the validity of using mixture models for estimating abundance in the presence of unobserved individual heterogeneity.

[1]Corresponding author:
Prof. Axel Munk, Institut für Mathematische Stochastik, Georg-August-University Göttingen, Maschmühlenweg 8-10, D-37073 Göttingen, Germany
email: munk@math.uni-goettingen.de
Fon: +49/551/39-13501/02/03, Fax: +49/551/39-13505

The aim of this paper is to examine the identifiability issue in more detail. In particular we prove identifiability *within* the mixture families that are most commonly used in this application. Thus, so long as the analyst is prepared to assume that the mixture distribution belongs to a certain family then identifiability is not a problem. Of course, if the analyst is prepared to make no assumptions about the distribution of probabilities then, as is well-known in the context of mixture models in general, Link's conclusion is correct.

2 Notation and preliminaries

Suppose that a closed population of unknown size N is sampled on T occasions. We assume that the number of captures X_i of animal i is distributed as Binomial $B(T, p_i)$, where p_i is the capture probability of this animal over T independent samples. We assume that the p_i's are distributed according to some distribution G on $[0, 1]$. This implies that the probability that an individual is sampled x-times is given by

$$\pi_G(x) = \binom{T}{x} \int_0^1 p^x (1-p)^{T-x} \, dG(p). \tag{1}$$

Let n be the number of animals which were captured at least once, i.e. for which $X_i > 0$. Let

$$f_x = \#\{i : X_i = x\}, \quad x = 1, \ldots, T.$$

As pointed out by Link (2003), the vector $(f_1, \ldots, f_T)$ is multinomially distributed with T cells, n repetitions and cell probabilities

$$\pi_G^c = \big(\pi_G^c(1), \ldots \pi_G^c(T)\big), \quad \pi_G^c(x) = \frac{\pi_G(x)}{1 - \pi_G(0)}, \quad x = 1, \ldots, T.$$

The probabilities π_G^c are simply the conditional probabilities of the mixture of binomial distributions given that $x \geq 1$. Note that only these conditional probabilities can be estimated from the observations f_x. Consequently the problem of establishing identifiability in this context differs from that of establishing identifiability of the mixing distribution G from the probabilities of a binomial mixture, π_G, in the classical mixture context (cf. Teicher, 1961, 1963; Lindsay, 1995). Here we need to investigate the identifiability of G from the conditional probabilities, π_G^c, given that $x \geq 1$. Note that once this identifiability is settled, G can be consistently estimated (within the given parametric family) by the maximum likelihood estimator $\hat{G}$, for example. Then $\pi_G(0)$ is consistently estimated by $\pi_{\hat{G}}(0)$, and N by

$$\hat{N} = \frac{n}{1 - \pi_{\hat{G}}(0)}.$$

By embedding the issue of identifiability in such capture-recapture models in the general context of identifiability of finite mixtures from the binomial distribution, it is immediately clear that G cannot be identified within the set of all distributions (for any fixed T), since this does not even hold for the complete (non-conditional) model (1) (cf. Teicher, 1961). However, in this note we show that *within* the commonly used parametric families, G is identifiable. Specifically, we show that within the class of finite mixtures with at most m components (cf. Pledger, 2000), G is identifiable if and only if $2m \leq T$. Furthermore, we give a general criterion for identifiability based on the moments of the mixing distribution. As particular

cases, this yields the identifiability of the class of beta distributions if $T \geq 3$, and of the class of uniform distributions on $[0, b]$, $b \leq 1$, if $T \geq 2$.
We stress that establishing identifiability in the context of this application is a more subtle problem than identifiability of binomial mixtures with fixed T, since only the conditional probabilities are available. Finally we remark that some of our results carry over to other distributions of the X_i's, such as the truncated geometric (cf. Norris and Pollock, 1996).

3 Theory and Examples

Definition 1. In the capture-recapture model (1) we shall call a family $\mathcal{G}$ of distributions on $[0, 1]$ *identifiable* if, for each $G \in \mathcal{G}$, the vector π_G^c uniquely determines G within the class $\mathcal{G}$, i.e. if for $G, H \in \mathcal{G}$,

$$\pi_G^c = \pi_H^c \Rightarrow G = H. \tag{2}$$

Lemma 1. *Let $(\pi_0, \ldots, \pi_T)$ and $(\rho_0, \ldots \rho_T)$ be two probability vectors on $\{0, \ldots, T\}$, and let π^c and ρ^c be the conditional probability vectors on $1, \ldots, T$, given that $x \geq 1$. Then*

$$\pi^c = \rho^c \Leftrightarrow \exists\ A > 0\ :\ \pi(x) = A\rho(x), \quad x = 1, \ldots, T.$$

It is known (e.g. Link, 2003) that for identifiability to be possible, 0 has to be excluded from the support of the distributions in $\mathcal{G}$. Indeed, let G be any distribution on $[0, 1]$ and consider $H = \lambda\delta_0 + (1 - \lambda)G$, $\lambda \in (0, 1)$. Since $\pi_{\delta_0}(0) = 1$, we have that $\pi_H(x) = (1 - \lambda)\pi_G(x)$ for $x = 1, \ldots, T$. From Lemma 1, it follows that $\pi_G^c = \pi_H^c$. Thus in the following we will concentrate on distributions G with support in $(0, 1]$.

Example 1 (*Finite mixtures*)**.** Consider the class of finite mixing distributions with m support points

$$\mathcal{G}_m = \Big\{G = \sum_{k=1}^{m} \lambda_k \delta_{p_k}, \quad \lambda_k \geq 0, \quad \sum_k \lambda_k = 1, \quad p_k \in (0, 1]\Big\}.$$

Theorem 1. *For $2m \leq T$ the class $\mathcal{G}_m$ is identifiable.*

Pledger (2000) used finite mixtures to model population heterogeneity. She observed that the condition $2m \leq T$ is necessary for identifiability. In fact, the class $\mathcal{G}_m$ has $2m - 1$ parameters, and these have to be identified from the $T - 1$ variable probabilities. Thus $2m \leq T$ is a necessary and sufficient condition. An inspection of the proof of Theorem 1 (cf. the Appendix) shows that the same arguments apply if the outcomes X_i follow a discrete distribution which, as a function of the parameter, is a Čebyšev system (cf. Karlin and Studden, 1966) with a joint zero outside the interval $(0, 1]$. An example is the truncated geometric distribution (for which $P(X_i = x) = p_i(1 - p_i)^x$, $1 \leq x \leq T$) used in Norris and Pollock (1996), to model population heterogeneity with behavorial response to capture.

We now turn to the case of continuous mixing distributions. Teicher (1961) observed that the probabilities (1) can be expressed in terms of the moments of the mixing distribution G. In fact, we have that

$$\pi_G(x) = \binom{T}{x} \sum_{k=x}^{T} (-1)^{k-x} \binom{T - x}{k - x} m_G(k), \quad x = 1, \ldots, T, \tag{3}$$

where $m_G(k) = \int_0^1 t^k \, dG(t)$ is the kth moment of G. For our problem this implies

Theorem 2. *For two distributions G, H on $(0,1]$, $\pi_G^c = \pi_H^c$ implies that there is an $A > 0$ such that*

$$m_G(x) = A\, m_H(x), \quad x = 1, \ldots, T. \tag{4}$$

Therefore if (4) does not hold for any two $G, H \in \mathcal{G}$, then $\mathcal{G}$ is identifiable.

Example 2. The beta distribution $B(p,q)$, $p,q > 0$ was used as a mixing distribution by Dorazio and Royle (2003). We show that this family is identifiable if $T \geq 3$. The xth moment is given by

$$m_{p,q}(x) = \frac{(p+x-1)\cdot \ldots \cdot p}{(p+q+x-1)\cdot \ldots \cdot (p+q)}.$$

¿From $m_{p,q}(x) = A m_{p',q'}(x)$, $x = 1,2,3$, and some $A > 0$, it follows that

$$\frac{(p+i)}{(p+q+i)} = \frac{(p'+i)}{(p'+q'+i)}, \quad i = 1,2.$$

Straightforward algebra now shows that $p = p'$ and $q = q'$.

Example 3. The uniform distribution on $(0,b]$ was considered as a mixing distribution in Pledger (2004). The first two moments are given by $m_b(1) = b/2$ and $m_b(2) = b^2/3$. From these expressions and Theorem 2 it is simple to see that for $T \geq 2$ the uniform distribution is identifiable.

Acknowledgement

Hajo Holzmann and Axel Munk gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft, Grant MU 1230/8–1.

References

Agresti, A. (1994). Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics* **50**, 494-500.

Borchers, D. L., Buckland, S. T. and Zucchini, W. (2002). *Estimating Animal Abundance: Closed Populations*. London: Springer.

Burnham, K.P. (1972). Estimation of population size in multinomial capture-recapture studies when capture probabilities vary among animals. Ph.D. Thesis, Oregon State University, Oregon.

Dorazio, R. M. and Royle, J. A. (2003). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics* **59**, 351-364.

Huggins, R. (2001). A note on the difficulties associated with the analysis of capture-recapture experiments with heterogeneous capture probabilities. *Stats. Probab. Lett.* **54**, 147–152.

Karlin, S. and Studden, W. J. (1966). *Tchebycheff systems: With applications in analysis and statistics.* New York: John Wiley & Sons.

Lindsay, B. (1995). *Mixture Models: Theory, Geometry, and Applications.* Hayward, CA: Institute for Mathematical Statistics.

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

Norris, J. L. and Pollock, K. H. (1995). A capture-recapture model with heterogeneity and behavioural

response. *Environmental and Ecological Statistics* **2**, 305-313.

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

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

Pledger, S. (2004). The performance of mixture models in heterogeneous closed population capture-recapture. *Preprint.*

Prautzsch, H., Boehm, W. and Paluszny, M. (2002). *Bézier and B-spline techniques.* New York: Springer.

Teicher, H. (1961). Identifiability of mixtures. *Ann. Math. Statist.* **32**, 244-248.

Teicher, H. (1963). Identifiability of finite mixtures. *Ann. Math. Statist.* **34**, 1265-1269.

Seber, G.A.F. (1982). *The estimation of animal abundance and related parameters.* 2nd edn. London: Charles Griffin.

Appendix

Proof of Lemma 1. For $x = 1, \ldots, T$,

$$\frac{\pi(x)}{1-\pi(0)} = \frac{\rho(x)}{1-\rho(0)} \Leftrightarrow \frac{\pi(x)}{\rho(x)} = \frac{1-\pi(0)}{1-\rho(0)} =: A.$$

□

First we prove the following lemma.

Lemma 2. *If*

$$\sum_{k=1}^{T} t_k \, p_k^x (1-p_k)^{T-x} = 0, \quad x = 1, \ldots, T,$$

for some $t_k \in \mathbb{R}$ *and* $p_k \in (0,1]$*, then it follows that* $t_1 = \ldots = t_T = 0$*.*

Proof. The polynomials $P_x(p) = p^x(1-p)^{T-x}$, $x = 1, \ldots, T$, are linearly independent because, except for the normalization, these are the Bernstein polynomials, which are known to be linearly independent, cf. Prautzsch et al., 2002. Therefore any linear combination has at most T roots. Since one of these always equals 0, there are at most $T-1$ roots within the interval $(0,1]$. Therefore for different $p_1, \ldots, p_T \in (0,1]$, if

$$\sum_{x=1}^{T} s_x p_k^x (1-p_k)^{T-x} = 0, \quad k = 1, \ldots, T,$$

it follows that $s_1 = \ldots = s_T = 0$. This implies that the matrix $(p_k^x(1-p_k)^{T-x})_{k,x=1,\ldots,T}$ has full rank. From this the statement of the lemma follows immediately. □

Proof of Theorem 1. Suppose that $G, H \in \mathcal{G}_m$ with $\pi_G^c = \pi_H^c$. From Lemma 1, there exists an $A > 0$ with

$$\sum_{k=1}^{m} \lambda_{k,G} \, p_{k,G}^x \, (1-p_{k,G})^{T-x} = A \sum_{k=1}^{m} \lambda_{k,H} \, p_{k,H}^x \, (1-p_{k,H})^{T-x},$$

$x = 1, \ldots, T$. Subtracting the r.h.s. from the l.h.s. and applying Lemma 2 implies that the support points of G and H coincide, and that $\lambda_{k,G} = A\lambda_{k,H}$ for these k (after a permutation). But since $\sum_k \lambda_{k,G} = \sum_k \lambda_{k,H} = 1$, $A = 1$. Therefore $G = H$. □

Proof of Theorem 2. In matrix form the identity (3) can be written as

$$(\pi_G(1), \dots \pi_G(T)) = M(\pi_H(1), \dots \pi_H(T)).$$

The matrix M is invertible, because it is upper triangular with nowhere vanishing diagonal. Therefore

$$\pi_G(x) = A\pi_H(x), \quad x = 1, \dots, T \Leftrightarrow m_G(x) = Am_H(x), \quad x = 1, \dots, T.$$

□