STATISTICA, anno LXXV, n. 1, 2015

LATENT CLASS RECAPTURE MODELS WITH FLEXIBLE BEHAVIOURAL RESPONSE

Alessio Farcomeni [1]
Dipartimento di Sanitá Pubblica e Malattie Infettive, Sapienza - Università di Roma, Roma, Italia

1. Introduction

The term capture-recapture denotes a class of experimental designs concerned with estimation of the size of a population. It has originally been employed in biometrics. The first applications regarded estimation of the number of residents in France (Laplace, 1783) and the size of fish populations in confined waters (Petersen, 1896). There are many possible ways to design a capture-recapture experiment. Loosely speaking, we focus in this paper on experiments in which as many subjects as possible are directly sampled from the population and identified without uncertainty. They are then released if necessary, and the operation is repeated $S > 1$ times. The binary capture history $\boldsymbol{y}_i$ of each subject, where $y_{ij} = 1$ if the i-th subject has been observed at the j-th occasion, can be used to infer on the number of subjects never captured or, equivalently and more interestingly, on the unknown population size N. There is an impressive literature on population size estimation, and we point the reader to the excellent reviews and overview of applications provided by Otis *et al.* (1978), Pollock (2000), Basu and Ebrahimi (2001), Chao (2001), Chao *et al.* (2003), Amstrup *et al.* (2003).

Population size estimates are obtained after appropriately modeling the capture histories, under certain assumptions on the data generating mechanisms. Assumptions include independence of capture histories at the subjects' level and that the population is closed, that is, no births, immigration, emigration and deaths occur during the sampling period. This assumption is often debated, but tenable when the time frame of data collection is sufficiently short.

Capture histories may be allowed to depend on subject-specific observed or unobserved covariates. In the first case we have M_o models, where the most general and difficult situation is that of individual covariates, which are missing for all subjects never captured. See for instance Royle (2009). In the second case we have M_h models where random effects are introduced to capture unobserved heterogeneity. Additionally, capture histories may depend on occasion-dependent

[1] Corresponding Author. E-mail: alessio.farcomeni@uniroma1.it

features (like weather conditions), making capture more likely or more unlikely uniformly for all subjects at each occasion. These are called M_t models, where occasion-specific capture probabilities are used. Finally, the probability of capture may depend on what happened during the previous capture occasions. Subjects may be more likely or less likely to be captured again after a capture, or a pattern of captures, depending on the experience. For instance, animals being fed during capture may become trap-happy, and looking forward to be captured and fed again. Models taking into account possible behavioural response to capture are tagged M_b.

These sources of heterogeneity may be combined, to obtain the most general possible model, tagged M_{hotb}. For reasons in our opinion linked with the applications involved, traditionally M_{htb} models were treated separately from M_{to} models. Farcomeni and Scacciatelli (2013) in a continuous time framework, and later Farcomeni (2015), were the first to consider the most general possible class of M_{hotb} models.

When dealing with capture-recapture data, the researcher directly faces the ubiquitous modeling dilemma between bias and variance of the estimates. More general models very closely fitting the data may provide highly variable population size estimates, while simpler models may lead to biased estimates. On the other hand, capture-histories may exhibit subtle dependency patterns leading even the most complex traditional approaches to be biased.

Traditional behavioural models for instance were rather rigid. In traditional M_b capture probabilities depended on past occasions only through a single update after the first capture event. The capture probability was afterwards constant regardless of the other capture events. This lead some authors to develop more flexible behavioural models which could better fit the data. Persistence models are proposed in Ramsey and Usner (2003), where a parameter controls the degree of attraction or repulsion of subjects towards identification. This effect is overlapped with possible persistence of the experimenters in returning to the same locations, from which the name of the model class. This overlap is of course not a problem given that behavioural effects are often merely nuisance parameters. Yang and Chao (2005) propose very interesting Markov chain models with *short-term* and *long-term* memory effects. Long-term memory effects are permanent changes in the capture-probability after the (first) capture, while short-term memory effects are soon forget and the initial propensity to capture is restored. Farcomeni (2011) proposed a completely general M_{tb} model where a completely arbitrary dependence among capture occasions is obtained. These do not even need to be ordered along a time horizon. The method is based on equality constraints for the conditional capture probabilities. For the M_{tb} models in Farcomeni (2011) closed form expressions can be obtained for the MLE of nuisance parameters, and a simple estimating equation for the population size. See also Alunni Fegatelli and Tardella (2013).

The approach of Farcomeni (2011) is extended in Farcomeni (2015) to include observed and unobserved heterogeneity after a logistic reparameterization of the capture probabilities. A class of M_{hotb} models in discrete time, where the *tb* part is completely general, is therefore obtained in Farcomeni (2015). Farcomeni

(2015) considers several possibilities for the distribution of random effects, but is mostly focused on continuously supported distributions. On the other hand, latent class models are very popular in capture-recapture given their flexibility. Latent class models can in fact be used to (i) identify subpopulations within the captured sample and (ii) approximate virtually any true underlying distribution, including continuous ones (in a completely different framework, see Bartolucci and Farcomeni (2009) for a simulation illustrating this, and the discussion of Bartolucci *et al.* (2014)).

The choice of a random effects distribution is rather delicate in capture-recapture, as usually there is substantial dependence of the final estimates on this choice. Additionally, a non-parametric MLE in which the distribution of random effects is unspecified is not possible (Link, 2003), unless one is prepared to consistently estimate a so-called meaningful lower bound for the population size (Farcomeni and Tardella, 2012). See also Farcomeni and Tardella (2010) for a Bayesian approach to M_{htb} estimation without assumptions on the underlying random effects distribution, Holzmann *et al.* (2006) for models based on minor assumptions (e.g., unimodality) on the random effects distribution, and Mao (2008) for one of the first definitions of estimable lower bound for the population size.

Latent class models were recommended as a general choice in many papers, including Pledger (2005) and Mao and You (2009). They are routinely employed in capture-recapture models in many fields, see for instance Coull and Agresti (1999), Bartolucci and Forcina (2006), Thandrayen and Wang (2010).

Farcomeni (2015) mentions latent class models and gives some guidelines on them, but an explicit definition of latent class M_{hotb} models and an explicit inferential procedure are not given. In this paper we will show how to fit M_{hotb} models proposed in Farcomeni (2015) under the relevant latent class assumption for the random effects. This is admittedly a special case of the completely general M_{hotb} model proposed in Farcomeni (2015). The inferential strategy will be along similar lines. We nevertheless believe that expliciting and detailing latent class M_{hotb} models is important as latent class models are widespread and generally recommended in population size estimation under heterogeneity, as argued above.

The rest of the paper is as follows: in the next section we revisit Farcomeni (2011) approach. In Section 3 we extend the approach to observed and unobserved heterogeneity based on latent class models. In Section 4 we show an efficient algorithm to obtain the MLE and estimate the population size.

2. Completely general M_{tb} models

Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iS})$, $i = 1, \ldots, N$, denote the binary capture history for the i-th subject. Without loss of generality, for $i = 1, \ldots, n$ we have that $\sum_j y_{ij} > 0$, that is, the first n subjects are captured at least once. On the other hand, for $i > n$, $\boldsymbol{y}_i = (0, \ldots, 0)$. The capture history of the $N - n$ subjects never captured is known and corresponds to a vector of zeros. On the other hand, we do not know how many subjects we have never sampled.

Let $p(\boldsymbol{y})$ denote the probability of a capture history $\boldsymbol{y}$. The joint distribution of binary capture indicators for the i-th subject, according to the chain rule, can be expressed as

$$p(\boldsymbol{y}) = p(Y_{i1} = y_{i1})p(Y_{i2} = y_{i2}|y_{i1}) \cdots p(Y_{iS} = y_{is}|y_{i,S-1}, \dots, y_{i1}).$$

Call now $p_1 = \Pr(Y_{i1} = 1)$, $p_j(a_{j-1}, \dots a_1) = p(Y_{ij} = 1|Y_{i,j-1} = a_{j-1}, \dots, Y_{i1} = a_1)$; for $a_j = \{0, 1\}$. These conditional capture probabilities are arranged in lexicographical order in the vector $\boldsymbol{p} = (p_1, p_2(0), p_2(1), p_3(00), \cdots, p_S(1, \dots, 1))$. In this paper we work with the conditional likelihood, that is, we condition on observed subjects. This leads to divide the likelihood by $\Pr(\sum_j Y_{ij} > 0)$. The conditional likelihood can be written as

$$\prod_{i=1}^{n} \frac{\prod_{j=1}^{S} p_j(Y_{i,j-1}, \dots, Y_1)^{Y_{ij}}(1 - p_j(Y_{i,j-1}, \dots, Y_1))^{1-Y_{ij}}}{1 - \prod_{j=1}^{S}(1 - p_j(0, \dots, 0))}, \tag{1}$$

where $p_1(\cdot) = p_1$ for notational convenience. In (1), the numerator gives $\Pr(Y_{i1}, \dots, Y_{iS})$ by definition, and the denominator $\Pr(\sum_j Y_{ij} > 0)$.

Let now $\boldsymbol{C}$ denote a matrix with exactly one -1 and one 1 per row, where all other entries are zeros. Equality constraints for the conditional capture probabilities can be expressed as $\boldsymbol{C}'p = 0$. The likelihood is as in (1), but only when $\boldsymbol{C}'p = 0$. It is otherwise $-\infty$.

It is shown in Farcomeni (2011) that all possible M_{tb} models are obtained by varying $\boldsymbol{C}$, and that for fixed $\boldsymbol{C}$ the MLE can be found with two simple estimating equations.

The role of the $\boldsymbol{C}$ matrix is that of selecting one among any possible M_{tb} model. This is achieved by constraining some conditional capture-history probabilities to be equal to each other. If for instance $p_2(0) = p_2(1)$, we have that captures at the second occasion are independent of captures at the first occasion. A $\boldsymbol{C}$ matrix must be always specified. When $\boldsymbol{C}$ is the empty matrix we have the saturated model based on $2^S - 1$ free parameters, and the MLE for the population size corresponds to n, the number of subjects observed at least once. More interesting estimates are obtained by constraining at least two parameters to be equal. The maximum number of rows of $\boldsymbol{C}$ is $2^S - 2$, where it is possible to obtain only one free parameter and therefore a model $M_\emptyset$, where probability of capture is homogeneous. If more than $2^S - 2$ rows are specified for $\boldsymbol{C}$, a redundant model specification is obtained. No issues arise for inference though, as some equalities would simply have been specified more than once. For any $\boldsymbol{C}$, a minimum of one and a maximum of $2^S - 1$ free parameters are obtained. See also Farcomeni (2011) and Farcomeni (2015) for more details.

The likelihood is always identifiable regardless of $\boldsymbol{C}$ (Farcomeni, 2015).

To fix the ideas, we give two examples of $\boldsymbol{C}$ matrices. Let $\boldsymbol{C} = \boldsymbol{D}_{2^S-2}$, where $\boldsymbol{D}_h = \{\boldsymbol{0}_{h-1,1}\ \boldsymbol{I}_{h-1}\} - \{\boldsymbol{I}_{h-1}\ \boldsymbol{0}_{h-1,1}\}$ is a matrix that produces first differences. Here $\boldsymbol{I}$ and $\boldsymbol{0}$ indicate identity and zero matrices of the specified size, and $\{A\ B\}$ indicates that the matrices A and B, having the same number of rows, have been column combined. If we assume $\boldsymbol{C}'p = 0$, we have that all parameters are equal to each other. This leads to the homogeneous model $M_\emptyset$. If we remove the first row from this matrix, p_1 is left completely free and we obtain an M_t model based on two parameters. Capture probabilities from the second occasion onwards are constrained to be equal to each other regardless of the previous capture history.

There are also ways to specify the model more simply, for instance by specifying directly one or more blocks of parameters that are assumed to be equal to each other.

3. Latent class M_{hotb} models

We now generalize the M_{tb} models introduced in the previous section to a latent class M_{hotb} model. Here the h stands for *unobserved heterogeneity*, which is accounted for through the latent class model. The o stands for *observed heterogeneity*, indicating the use of covariates. All possible sources of inhomogeneity in subject and occasion specific capture probabilities will therefore be taken into account.

Observed and unobserved heterogeneity is introduced after a logit reparameterization. See Huggins (1989), Alho (1990), Coull and Agresti (1999, 2000). The parameterization for the M_{tb} class is as follows:

$$\log\left(\frac{p_j(a_{j-1},\dots,a_1)}{1-p_j(a_{j-1},\dots,a_1)}\right)=\beta_{ja_1,\dots,a_{j-1}}. \tag{2}$$

It shall be noted that equality constraints are invariant to this parameterization, as $\boldsymbol{C}'\boldsymbol{p}=0$ if and only if $\boldsymbol{C}'\boldsymbol{\beta}=0$.

Let $\boldsymbol{X}_{ij}$ denote a subject- and time-specific vector of covariates for the i-th subject. The conditional probabilities also depend on a subject-specific parameter θ_i, which summarizes subject-specific unobserved heterogeneity. Our final model is then

$$\log\left(\frac{p_j(a_{j-1},\dots,a_1,\theta,\boldsymbol{X})}{1-p_j(a_{j-1},\dots,a_1,\theta,\boldsymbol{X})}\right)=\beta_{ja_1,\dots,a_{j-1}}+\theta+\boldsymbol{\gamma}'\boldsymbol{X}, \tag{3}$$

for $j=1,\dots,S$. In (3) above we have an M_{hotb} model as $\boldsymbol{\beta}$ parameters summarize the occasion-specific and behavioural effects, θ summarizes unobserved heterogeneity, and $\boldsymbol{\gamma}$ describes the effects of covariates $\boldsymbol{X}$.

It is important to underline here that $\boldsymbol{X}_{ij}$ is not observed for subjects never captured, that is, for $i>n$. We will be able to estimate parameters by relying on the so called conditional likelihood, that is, by restricting to subjects observed at least once.

Equality constraints are still specified as $\boldsymbol{C}'\boldsymbol{\beta}=0$, but now have a slight different meaning as they lead to equality of capture probabilities only for subjects sharing the same configuration of observed and unobserved covariates.

We assume the random effects arise from a latent class model based on k support points $\xi_1,\dots,\xi_k$, with probabilities $\pi_1,\dots,\pi_k$, so that

$$\Pr(\theta=\xi_j)=\pi_j$$

for $j=1,\dots,k$; and $\sum_j \pi_j=1$. For identifiability reasons we must impose a constraint on the support points vector. After some algebra it is straightforward indeed to see that without constraints adding a constant to each ξ and substracting the same from $\beta_j(a_1,\dots,a_{j-1})$ leads to exactly the same value for the likelihood. A simple and interpretable constraint is that $\sum_j \xi_j=0$, which we will assume throughout. We will discuss in the next section how to estimate $\boldsymbol{\xi}$ under the sum-to-zero constraint. Other possible constraints include a corner point parameterization where $\xi_l = 0$ for some l. In both cases, it is not possible to add a constant to c to each ξ, as it would violate the constrain. In fact, $\sum_j \xi_j$ or xi_l would be equal to c, for $c \neq 0$.

It shall be noted here that the number of latent classes should be specified in advance. The usual and recommended route is to repeatedly fit the model for different choices of k, and then select the best k by trading-off bias (which is larger for smaller k) and variance (which is larger for larger k). An automatic and simple way to do so is by optimizing over k some information criterion, like Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). See Akaike (1973), Anderson *et al.* (1994), Burnham *et al.* (1995), Schwarz (1978). AIC is more appropriate for small sample sizes and it is optimal from a predictive point of view, BIC is model consistent.

4. Inference through the EM algorithm

As mentioned in the previous section, we work with the *conditional* likelihood. It has been shown in Sanathanan (1972) and Fewster and Jupp (2009) that under parametric assumptions for the random effects this is equivalent to maximization of the complete likelihood. More importantly, parameter estimates are consistent and asymptotically normal at the usual rate.

The conditional likelihood is simply obtained by dividing the probability of each observed capture history by the probability that the subject is indeed observed. That is, the individual likelihood contribution of each observed subject $\Pr(\boldsymbol{y}_i)$ is divided by probability of recording him/her/it $\Pr(\sum_j y_{ij} > 0)$; therefore obtaining a conditional capture probability. The resulting conditional likelihood does not depend on N or on the unobserved covariates.

An additional problem with our model formulation is that the latent class of each subject is not observed. Therefore, we must not only *condition* the likelihood on the occurrence of sampling, but also *complete* it with the unobserved latent classes. The resulting *complete conditional likelihood* (Farcomeni and Scacciatelli, 2013; Farcomeni, 2015) will be used to obtain an expectation maximization (EM) algorithm which will lead to a maximum of the conditional likelihood.

The complete conditional log-likelihood is defined when $\boldsymbol{C}'\boldsymbol{\beta} = 0$ and is given by the following expression:

$$
\begin{aligned}
l_c(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) \quad = & \quad \sum_{i=1}^{n}\sum_{j=1}^{S}\sum_{h=1}^{k} Z_{ih} Y_{ij} \log(p_j(Y_{i,j-1}, \dots, Y_{i1}, \xi_h, \boldsymbol{X}_i)) + \\
+ & \quad (1 - Y_{ij}) \log(1 - p_j(Y_{i,j-1}, \dots, Y_{i1}, \xi_h, \boldsymbol{X}_i)) + \\
- & \quad \sum_{h=1}^{k}\sum_{i=1}^{n} Z_{ih} \log(1 - \prod_{j=1}^{S}(1 - p_j(0, \dots, 0, \xi_h, \boldsymbol{X}_i))) \\
+ & \quad \sum_{h=1}^{k}\sum_{i=1}^{n} Z_{ih} \log(\pi_h),
\end{aligned}
\tag{4}
$$

where $Z_{ih} = 1$ if the i-th subject is in class h, and zero otherwise; and with the convention that $p_1(\cdot, \theta_i, \boldsymbol{X}_i) = p_1(\theta_i, \boldsymbol{X}_i)$.

It is now a matter of setting up a constrained EM algorithm to obtain the MLE. The EM algorithm alternates two steps until convergence. At the E step the expected value of (4) with respect to the current posterior distribution of the random effects is obtained. At the M step the expected value above is maximized under the constraints with respect to the parameters.

As usual, the EM algorithm is guaranteed to converge only to a local optimum of the conditional likelihood function. The EM algorithm is then repeatedly run from different initial solutions in order to increase the possibilities of finding the global optimum.

Standard errors and confidence intervals for the parameters, conditionally on n, may be obtained through a nonparametric bootstrap procedure. The subjects observed at least once are resampled with replacement and the EM algorithm is used on these fictitious samples. The standard deviation of the set of resulting estimates is a good approximation to the conditional standard errors. Unconditional estimates involve only an additional term, described in Böhning (2008).

4.1. E step

It is straightforward to check that the complete conditional likelihood above depends *linearly* on missing values Z_{ih}. Therefore, we can obtain the expected value of (4) simply by plug-in of the posterior expectation of Z_{ih}. This is as follows:

$$\begin{aligned} \tilde{z}_{ih} = \Pr(Z_{ih} = 1|\boldsymbol{Y}) &\propto \Pr(Y|Z_{ih} = 1)\Pr(Z_{ih} = 1) = \\ &= \pi_h \prod_{j=1}^{S} p_j(Y_{i,j-1}, \dots, Y_{i1}, \xi_h, \boldsymbol{X}_i)^{Y_{ij}} \\ &\quad (1 - p_j(Y_{i,j-1}, \dots, Y_{i1}, \xi_h, \boldsymbol{X}_i))^{1-Y_{ij}} \\ &\quad (1 - \prod_{j=1}^{S}(1 - p_j(0, \dots, 0, \xi_h, \boldsymbol{X}_i)))^{-1}, \end{aligned} \tag{5}$$

where the normalizing constant is given by the sum over h of the right hand side of the expression above. It is important to underline that (5) is evaluated conditionally on the current parameter estimates.

4.2. M step

At the M step we maximize the expected complete conditional likelihood under the constraint $\boldsymbol{C}'\boldsymbol{\beta} = 0$. It is straightforward to check that a closed form expression is available for $\hat{\pi}_h$, that is,

$$\hat{\pi}_h = \frac{\sum_i \tilde{z}_{ih}}{\sum_i \sum_h \tilde{z}_{ih}}.$$

Maximization of the rest of the complete conditional likelihood is slightly more cumbersome, and we proceed as in Farcomeni (2015) through the Aitchison and Silvey (1958) (AS) algorithm.

The AS algorithm recognizes that optimization of the expected complete conditional likelihood is equivalent to the system of non-linear equations

$$\begin{cases} \boldsymbol{s}(\boldsymbol{\eta}|\boldsymbol{\eta}') + \boldsymbol{C}\boldsymbol{\lambda} = 0 \\ \boldsymbol{C}'\boldsymbol{\beta} = 0, \end{cases}$$

where $\boldsymbol{\eta} = (\boldsymbol{\beta}\ \boldsymbol{\gamma}\ \boldsymbol{\xi})$, $\boldsymbol{s}(\boldsymbol{\eta}|\boldsymbol{\eta}')$ denotes the gradient of the expected complete likelihood with respect to $\boldsymbol{\eta}$, $\boldsymbol{\lambda}$ is a vector of unknown Lagrange multipliers, and $\boldsymbol{s}(\cdot|\cdot)$ denotes the score of the expected complete likelihood. The AS algorithm proceeds by substituting $\boldsymbol{s}(\boldsymbol{\eta}|\boldsymbol{\eta}')$ with a first order linear approximation based on the Hessian $-\boldsymbol{I}(\boldsymbol{\eta}|\boldsymbol{\eta}')$, that is,

$$\boldsymbol{s}(\boldsymbol{\eta}|\boldsymbol{\eta}') \approx \boldsymbol{s}(\boldsymbol{\eta}_t|\boldsymbol{\eta}') - (\boldsymbol{\eta} - \boldsymbol{\eta}_t)'\boldsymbol{I}(\boldsymbol{\eta}_t|\boldsymbol{\eta}'),$$

where $\boldsymbol{\eta}_t$ is the value of $\boldsymbol{\eta}$ at the current iteration of the AS algorithm. The expression for $\boldsymbol{s}(\boldsymbol{\eta}|\boldsymbol{\eta}')$ is given in appendix, while $\boldsymbol{I}(\boldsymbol{\eta}_t|\boldsymbol{\eta}')$ is obtained as minus the numerical first derivative of $\boldsymbol{s}(\boldsymbol{\eta}|\boldsymbol{\eta}')$.

We now augment $\boldsymbol{C}$ to obtain $\boldsymbol{D}$, so that the constraint $\boldsymbol{C}'\boldsymbol{\beta} = 0$ is equivalent to $\boldsymbol{D}'\boldsymbol{\eta} = 0$. The augmented matrix $\boldsymbol{D}$ will contain rows of zeros in correspondence of the unconstrained slopes $\boldsymbol{\gamma}$. The sum to zero constraint $\sum_h \xi_h = 0$ is simply and directly obtained with a single column of $\boldsymbol{D}$ being $1/k$ in correspondence of the entries of $\boldsymbol{\xi}$ within $\boldsymbol{\eta}$, and zero otherwise.

Suppose now $\boldsymbol{D}'\boldsymbol{\eta}_t = 0$, e.g., that the algorithm is initialized from a set of parameters satisfying the constraints. Note that a feasible set of parameters can always be obtained by initializing $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ to the zero vector and π_h to $1/k$.

The approximated system of non-linear equations is exactly solved by the updating rule

$$\begin{aligned} \boldsymbol{\eta}_{t+1} &= \boldsymbol{\eta}_t + \boldsymbol{I}(\boldsymbol{\eta}_t|\boldsymbol{\eta}')^{-1}\boldsymbol{s}(\boldsymbol{\eta}_t|\boldsymbol{\eta}') + \\ &\quad -\boldsymbol{I}(\boldsymbol{\eta}_t|\boldsymbol{\eta}')^{-1}\boldsymbol{D}(\boldsymbol{D}'\boldsymbol{I}(\boldsymbol{\eta}_t|\boldsymbol{\eta}')^{-1}\boldsymbol{D})^{-1}(\boldsymbol{D}'\boldsymbol{I}(\boldsymbol{\eta}_t|\boldsymbol{\eta}')^{-1}\boldsymbol{s}(\boldsymbol{\eta}_t|\boldsymbol{\eta}')). \end{aligned} \tag{6}$$

The updating rule, which is similar to the Newton-Raphson one, is then iterated until convergence to obtain the current update of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$. The constraint will be exactly satisfied for $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$. See also Evans and Forcina (2013) for improvements and more details on the AS algorithm.

4.3. Population size estimates

An Horvitz-Thompson estimator for the population size is given by

$$\hat{N} = \sum_{i=1}^{n} \left(1 - \prod_{j=1}^{S} (1 - p_j(0, \ldots, 0, \hat{\xi}_i, \boldsymbol{X}_i)) \right)^{-1}, \tag{7}$$

where $p_j(\cdot)$ is obtained after plug-in the MLE for all parameters and $\hat{\xi}_i$ is the empirical Bayes estimate of the latent location for subject i. More precisely,

$$\hat{\xi}_i = \sum_{h=1}^{k} \xi_h \tilde{z}_{ih},$$

where the posterior $\tilde{z}_{ih}$ is evaluated at the MLE.

5. Conclusions

We have shown in detail a special case of the M_{hotb} class of models of Farcomeni (2015). Farcomeni (2015) introduces the model class based on a general (parametric) form for the mixing distribution, giving some guidelines on latent class models but focusing mostly on continuously supported random effects. In this paper we have explicited and discussed in details the subtleties (e.g., why and how to impose constraints on $\boldsymbol{\xi}$) linked with latent class M_{hotb} models.

The class of models proposed is very large and provides extreme flexibility for even the most complex capture-recapture experiments. It has proved very useful in applications in rather different fields, see for instance the examples developed in Farcomeni and Scacciatelli (2013) and Farcomeni (2015). In Farcomeni (2015) it is argued, through examples and extensive simulations, that time and behavioural heterogeneity can be rather complex in common ecological applications, and that specifying an overly simplified model can lead to substantial bias. The proposed approach is flexible enough to allow the researcher to capture heterogeneity. A form of misspecification that is particularly problematic, and not investigated in Farcomeni (2015), is that arising from the mixing distribution. In this paper we have used latent class models, which are somehow more flexible than continuous mixing distributions.

The are many possibilities for further work. First of all, standard errors are not directly available from the EM algorithm output. We have used resampling so far, but they could also possibly be obtained by combining results from Aitchison and Silvey (1958) and Oakes (1999), as in Bartolucci and Farcomeni (2015).

Secondly, data collection is often complex and possibly imprecise with capture-recapture, hence for M_o models the covariate matrices may include outliers. A possibility for further work includes providing *robust* inference for the M_{hotb} model class (see e.g. Farcomeni and Ventura (2012); Farcomeni and Greco (2015)).

Third, we rely on the conditional likelihood. In presence of unobserved heterogeneity, the complete likelihood may have advantages (Farcomeni and Tardella, 2012). We could work with the complete likelihood via data augmentation schemes as in Royle (2009).

Finally, the approach could be extended to the more difficult situation of *open* population models where the population size changes during the observation period. This is straightforward and involves only multiplying the likelihood by the appropriate individual survival and birth probabilities. Adjustments to the EM algorithm are though really substantial, and we leave them for further work.

Acknowledgements

This work was supported in part by a grant from Sapienza - University of Rome. The author is grateful to two referees for stimulating comments.

Appendix

A. Score of the expected complete log-likelihood

The score involved in (6) at the M step of the algorithm is given in this section. In the following we denote with $Q(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\xi}')$ the expected complete likelihood obtained after plug-in of $\tilde{z}_{ih}$. $\boldsymbol{\eta} = (\boldsymbol{\beta}\ \boldsymbol{\gamma}\ \boldsymbol{\xi})$ denotes the parameter vector to be optimized upon, while $\boldsymbol{\eta}'$ denotes the current parameter estimates which are used to compute $\tilde{z}_{ih}$.

Let $\boldsymbol{a} \neq 0$. For $j = 1, \ldots, S$;

$$\frac{\partial Q(\boldsymbol{\eta}|\boldsymbol{\eta}')}{\partial \beta_{j\boldsymbol{a}}} = \sum_{i=1}^{n} I(Y_{i1} = a_1, \ldots, Y_{i,j-1} = a_{j-1}) \left(Y_{ij} - \sum_h p_j(\boldsymbol{a}) \tilde{z}_{ih} \right), \tag{8}$$

where $I(\cdot)$ is the indicator function.

Let $\mathbf{0}$ denote a vector of zeros of the opportune size. We have

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\eta}|\boldsymbol{\eta}')}{\partial \beta_{j\mathbf{0}}} &= \sum_{i=1}^{n} I(Y_{i1} = 0, \ldots, Y_{i,j-1} = 0) \left(Y_{ij} - \sum_{h=1}^{k} p_j(\mathbf{0}) \tilde{z}_{ih} \right) + \\ &- \sum_{i=1}^{n} \sum_{h=1}^{k} \frac{p_j(\mathbf{0}) \prod_{h=1}^{S} (1 + \exp(\beta_{h\mathbf{0}} + \xi_h + \boldsymbol{\gamma}' \boldsymbol{X}_i))^{-1}}{1 - \prod_{h=1}^{S} (1 + \exp(\beta_{h\mathbf{0}} + \xi_h + \boldsymbol{\gamma}' \boldsymbol{X}_i))^{-1}} \tilde{z}_{ih} \end{aligned} \tag{9}$$

After some algebra it can also be seen that

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\eta}|\boldsymbol{\eta}')}{\partial \gamma_h} &= \sum_{l=1}^{k} \sum_{i=1}^{n} X_{ih} \sum_{j=1}^{S} \left(Y_{ij} - \int p_j(Y_{i,j-1}, \ldots, Y_{i1}) \tilde{z}_{il} \right) + \\ &- \sum_{i=1}^{n} X_{ih} \sum_{l=1}^{k} \left(\sum_{j=1}^{S} p_j(\mathbf{0}) \right) \frac{\prod_{j=1}^{S} (1 + \exp(\beta_{j\mathbf{0}} + \xi_l + \boldsymbol{\gamma}' \boldsymbol{X}_i))^{-1}}{1 - \prod_{j=1}^{S} (1 + \exp(\beta_{j\mathbf{0}} + \xi_l + \boldsymbol{\gamma}' \boldsymbol{X}_i))^{-1}} \tilde{z}_{il} \end{aligned} \tag{10}$$

Finally,

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\eta}|\boldsymbol{\eta}')}{\partial \xi_h} &= \sum_{i=1}^{n} \sum_{j=1}^{S} (Y_{ij} - \xi_h p_j(Y_{i,j-1}, \ldots, Y_{i1}, \xi_h, \boldsymbol{X}_i)) + \\ &- \sum_{i=1}^{n} \left(\sum_{j=1}^{S} p_j(\mathbf{0}, \xi_h) \right) \frac{\prod_{j=1}^{S} (1 + \exp(\beta_{j\mathbf{0}} + \xi_h + \boldsymbol{\gamma}' \boldsymbol{X}_i))^{-1}}{1 - \prod_{j=1}^{S} (1 + \exp(\beta_{j\mathbf{0}} + \xi_h + \boldsymbol{\gamma}' \boldsymbol{X}_i))^{-1}} \tilde{z}_{ih} \end{aligned} \tag{11}$$

References

J. Aitchison, S. D. Silvey (1958). *Maximum-likelihood estimation of parameters subject to restraints.* Annals of Mathematical Statistics, 29, pp. 813–828.

H. Akaike (1973). *Information theory as an extension of the maximum likelihood principle.* In B. N. Petrov, C. F. (eds.), *Second International symposium on information theory.* Akademiai Kiado, Budapest, pp. 267–281.

J. M. Alho (1990). *Logistic regression in capture-recapture models.* Biometrics, 46, pp. 623–635.

D. Alunni Fegatelli, L. Tardella (2013). *Improved inference on capture recapture models with behavioural effects.* Statistical Methods & Applications, 22, pp. 45–66.

S. C. Amstrup, T. L. McDonald, B. F. J. Manly (eds.) (2003). *Handbook of Capture- Recapture Analysis.* John Wiley, London.

D. R. Anderson, K. P. Burnham, G. C. White (1994). *Aic model selection in overdispersed capture-recapture data.* Ecology, 75, pp. 1780–1793.

F. Bartolucci, A. Farcomeni (2009). *A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure.* Journal of the American Statistical Association, 104, pp. 816–831.

F. Bartolucci, A. Farcomeni (2015). *Information matrix for hidden Markov models with covariates.* Statistics and Computing, 25, pp. 515–526.

F. Bartolucci, A. Farcomeni, F. Pennoni (2014). *Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates (with discussion).* TEST, 23, pp. 433–486.

F. Bartolucci, A. Forcina (2006). *A class of latent marginal models for capture-recapture data with continuous covariates.* Journal of the American Statistical Association, 101, pp. 786–794.

S. Basu, N. Ebrahimi (2001). *Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence.* Biometrika, 88, pp. 269–279.

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

K. P. Burnham, G. C. White, D. R. Anderson (1995). *Model selection strategy in the analysis of capture-recapture data.* Biometrics, 51, pp. 888–898.

A. Chao (2001). *An overview of closed capture-recapture models.* Journal of Agricultural, Biological and Environmental Statistics, 6, no. 2, pp. 158–175.

A. Chao, P. K. Tsay, L. S.-H., W.-Y. Shau, D.-Y. Chao (2003). *Tutorial in biostatistics: the applications of capture-recapture models to epidemiological data.* Statistics in Medicine, 20, pp. 3123–3157.

B. A. Coull, A. Agresti (1999). *The use of mixed logit models to reflect heterogeneity in capture-recapture studies.* Biometrics, 55, pp. 294–301.

B. A. Coull, A. Agresti (2000). *Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution.* Biometrics, 56, pp. 73–80.

R. J. Evans, A. Forcina (2013). *Two algorithms for fitting constrained marginal models*. Computational statistics & Data analysis, 66, pp. 1–7.

A. Farcomeni (2011). *Recapture models under equality constraints for the conditional capture probabilities*. Biometrika, 98, pp. 237–242.

A. Farcomeni (2015). *A general class of recapture models based on the conditional capture probabilities*. Biometrics, p. available online.

A. Farcomeni, L. Greco (2015). *Robust Methods for Data Reduction*. Chapman & Hall/CRC Press.

A. Farcomeni, D. Scacciatelli (2013). *Heterogeneity and behavioural response in continuous time capture-recapture, with application to street cannabis use in Italy*. Annals of Applied Statistics, 7, pp. 2293–2314.

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

A. Farcomeni, L. Tardella (2012). *Identifiability and inferential issues in capture-recapture experiments with heterogeneous detection probabilities*. Electronic Journal of Statistics, 6, pp. 2602–2626.

A. Farcomeni, L. Ventura (2012). *An overview of robust methods in medical research*. Statistical Methods in Medical Research, 21, pp. 111–133.

R. Fewster, P. Jupp (2009). *Inference on population size in binomial detectability models*. Biometrika, 96, pp. 805–820.

H. Holzmann, A. Munk, W. Zucchini (2006). *On identifiability in capture-recapture models*. Biometrics, 62, no. 3, pp. 934–936.

R. Huggins (1989). *On the statistical analysis of capture experiments*. Biometrika, 76, pp. 133–140.

P. Laplace (1783). *Sur les naissances, les marriages, et les morts*. Paris.

W. A. Link (2003). *Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities*. Biometrics, 59, no. 4, pp. 1123–1130.

C.-X. Mao (2008). *On the nonidentifiability of population sizes*. Biometrics, 64, no. 3, pp. 977–979.

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

D. Oakes (1999). *Direct calculation of the information matrix via the EM algorithm*. Journal of the Royal Statistical Society (Series B), 61, pp. 479–482.

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

C. G. Petersen (1896). *The yearly immigration of young plaice into the Limfjord from the German sea.* Tech. rep., Danish Biological Station.

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

K. H. Pollock (2000). *Capture-recapture models.* Journal of the American Statistical Association, 95, pp. 293–296.

F. Ramsey, D. Usner (2003). *Persistence and heterogeneity in habitat association studies using radio tracking.* Biometrics, 59, pp. 331–339.

J. A. Royle (2009). *Analysis of capture-recapture models with individual covariates using data augmentation.* Biometrics, 65, pp. 267–276.

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

G. Schwarz (1978). *Estimating the dimension of a model.* Annals of Statistics, 6, pp. 461–464.

J. Thandrayen, Y. Wang (2010). *Capture-recapture analysis with a latent class model allowing for local dependence and observed heterogeneity.* Biometrical Journal, 52.

H.-C. Yang, A. Chao (2005). *Modeling animals' behavioral response by Markov chain models for capture-recapture experiments.* Biometrics, 61, pp. 1010–1017.

Summary

We propose a class of models for population size estimation in capture-recapture studies, allowing for flexible behavioural and time response, observed heterogeneity and unobserved heterogeneity. The latter is taken into account by means of discrete random variables. The conditional likelihood is maximized through an efficient EM based on the Aitchinson-Silvey algorithm.

Keywords: Capture history; Equality constraints; Population size