Biometrics 72, 116–124
March 2016

DOI: 10.1111/biom.12375

A General Class of Recapture Models Based on the Conditional Capture Probabilities

A. Farcomeni*

Department of Public Health and Infectious Diseases, Sapienza - University of Rome
**email:* alessio.farcomeni@uniroma1.it

Summary. We propose an M_{hotb} model for population size estimation in capture-recapture studies. The *tb* part is based on equality constraints for the conditional capture probabilities, leading to an extremely rich model class. Observed and unobserved heterogeneity are dealt with by means of a logistic parameterization. In order to explore the model class, we introduce a penalized version of the likelihood. The conditional likelihood and penalized conditional likelihood are maximized by means of efficient EM algorithms. Simulations and two real data examples illustrate the approach.

Key words: Aitchinson-Silvey algorithm; Capture history; Equality constraints; Heterogeneity; Population size.

1. Introduction

Capture-recapture is concerned with estimation of the size of a closed population based on the capture history over repeated occasions. Capture probabilities may depend on four elements: the specific capture occasion (M_t models), the previous capture occasions (M_b models), observed heterogeneity (that is, covariates, called M_o models in (Farcomeni and Scacciatelli, 2013)) and unobserved heterogeneity (M_h models). These sources of heterogeneity may be combined, to obtain the most general possible model, tagged M_{hotb}. See also Otis et al. (1978). Traditional behavioral models let capture probabilities depend on past occasions by updating them after the first capture event. There are generalizations, see, for instance, Ramsey and Usner (2003) and Yang and Chao (2005). A completely general M_{tb} model is proposed in Farcomeni (2011), 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, and surprisingly enough there are closed form expressions for the MLE of nuisance parameters and a simple estimating equation for the MLE of the population size. The approach of Farcomeni (2011) has anyway two strong limitations: first, there is no way of including observed or unobserved heterogeneity; secondly, the set of possible constraints is large, but there are no suggestions in Farcomeni (2011) on how to explore the model class. The current practice is to specify a small set of candidate models and compare them using the Akaike Information Criterion (AIC) (Akaike, 1973; Anderson et al., 1994; Burnham et al., 1995). In this article we try to overcome these limitations. First, we generalize the conditional equality models in Farcomeni (2011) to include observed and unobserved heterogeneity. We deal with the most general case of individual covariates (e.g., Royle (2009)). The inferential approach is not straightforward anymore, but is anyway based on a general and very efficient algorithm for constrained inference in categorical data analysis. We then describe an approach for selecting the set of constraints. `R` code to reproduce our methods is available as supplementary material.

The rest of the article is as follows: in the next section we revisit Farcomeni (2011) approach. In Section 3 we extend the approach to observed and unobserved heterogeneity. In Section 4 we discuss how to select the best model within the very rich class obtained. We illustrate with a simulation study in Section 5 and an application in Section 6. Some concluding remarks are given in Section 7.

2. Set Up

Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iS})$, $i = 1, \ldots, N$, denote the binary capture history for the i-th subject, where we have $S > 1$ capture occasions and an unknown population size N. We observe capture histories of the n subjects for which $\sum_j y_{ij} > 0$. Let $p(\boldsymbol{y})$ denote the probability of a capture history $\boldsymbol{y}$. A fully general parameterization is given by the chain rule, so that

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

Call now $p_1 = \Pr(Y_{i1} = 1)$, $p_j(a_{j-1}, \ldots a_1) = p(Y_{ij} = 1|Y_{i,j-1} = a_{j-1}, \ldots, Y_{i1} = a_1)$; for $a_j = \{0, 1\}$. These conditional capture probabilities are arranged in lexicographical order in a vector $\boldsymbol{p} = (p_1, p_2(0), p_2(1), p_3(0,0), \cdots, p_S(1, \ldots, 1))$. The saturated model is based on these $2^S - 1$ parameters and leads to the trivial MLE $\hat{N} = n$. Let now $\boldsymbol{C}$ denote a binary matrix of contrasts, so that equality constraints can be expressed as $\boldsymbol{C}'p = 0$. 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. For instance, we could specify $\boldsymbol{C} = \boldsymbol{D}_{2^S-2}$, 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

© 2015, The International Biometric Society

first differences. Here $\boldsymbol{I}$ and $\boldsymbol{0}$ indicate identity and zero matrices of the specified size. In this way we would constrain all β parameters to be equal to each other, so to obtain a simple M_0 model. If we set

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{1}_{2^{s-1}-1,1} & & \boldsymbol{I}_{2^{s-1}-1} \otimes (0,-1)' & \\ \boldsymbol{0}_{2^{s-1}-2,1} & \boldsymbol{1}_{2^{s-1}-2,1} & \boldsymbol{I}_{2^{s-1}-1,-1} \otimes (0,-1)' & \boldsymbol{0}_{2^{s-1}-2,1} \end{pmatrix}, \tag{1}$$

where $\boldsymbol{1}$ indicates a matrix of ones, we would obtain a first-order Markovian structure as in Yang and Chao (2005). We term M_{c1} the resulting model. The innovation with respect to Yang and Chao (2005) is that we may possibly include also the *ho* parts in the following. Note that the maximum number of parameters is equal to the number of cells of the contingency table, and therefore the model is always identifiable regardless of $\boldsymbol{C}$. Further, given usual regularity conditions the MLE is asymptotically consistent and asymptotically normal under the model (Sanathanan, 1972).

3. A General M_{hotb} Model

In this section we obtain a fully general M_{hotb} model. We begin by introducing a mixed-effects logit reparameterization (Coull and Agresti, 1999, 2000), which will be convenient for inclusion of categorical and continuous predictors.

Let us specify the following parameterization:

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

The logit transformation simply maps the 2^S-1 parameters $\boldsymbol{p}$ to the 2^S-1 parameters $\boldsymbol{\beta} \in \mathcal{R}^{2^S-1}$. Equality constraints are imposed similarly, as $\boldsymbol{C}'\boldsymbol{p}=0$ if and only if $\boldsymbol{C}'\boldsymbol{\beta}=0$.

Let $\boldsymbol{X}_{ik}$ denote a subject- and time-specific vector of covariates for the i-th subject. One way of dealing with observed heterogeneity is to rely on a logistic reparameterization and condition the likelihood on subjects observed at least once, so that there are no missing covariates. See Huggins (1989) and Alho (1990). Furthermore, we assume that the conditional probabilities also depend on a subject-specific parameter θ_i, which summarizes subject-specific unobserved heterogeneity. Our final model can be expressed as

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

for $j=1,\ldots,S$. We make the following assumptions in the model above: (i) *additivity*, that is, the effects of observed and unobserved heterogeneity are additive on the log-scale (also known as proportionality of the odds) and (ii) *slope homogeneity*, that is, that the effects of the covariates $\boldsymbol{\gamma}$ do not depend on the specific capture occasion, past capture history, and unobserved heterogeneity. More explicitly, the h, o, and tb effects do not interact. A consequence of slope homogeneity, which can be easily relaxed to some extent if needed, is that all occasion-specific and behavioral effects are captured by the parameter $\boldsymbol{\beta}$. The model is completed by equality constraints on the *adjusted* conditional capture probabilities, in the form of an opportune constraint $\boldsymbol{C}'\boldsymbol{\beta}=0$; and with an assumption for the random effect of the kind $\theta \sim F(\boldsymbol{\alpha})$. Two subjects with the same covariate configuration X and the same subject-specific parameter θ might then have the same capture-history probabilities even with different capture histories. This happens if the capture histories belong to the same partition implied by the constraint $\boldsymbol{C}'\boldsymbol{\beta}=0$. By varying the matrix $\boldsymbol{C}$ and the distributional assumption on θ, a huge number of previously known and completely new M_{hotb} models can be obtained. The issue of model selection in this vast model class will be dealt with below.

There are many parametric choices for the mixing distribution F, including Gaussian, Student's T, univariate symmetric Laplace, logit-Beta, latent class. Non-parametric assumptions may not be pursued due to identifiability issues given that we work with the conditional likelihood (Link, 2003; Farcomeni and Tardella, 2012). It is then recommended that as usual few options are specified and compared through information criteria like AIC.

3.1. *Inference through the EM Algorithm*

The covariate configuration of the unobserved subjects, whose capture history is summarized by $Y_{ij}=0$ for $j=1,\ldots,S$, is not known. We work with the *conditional* likelihood, that is, we condition each capture history probability to the event that the subject is captured at least once in the S occasions. The resulting likelihood expression does not depend on N or on the unobserved covariates. Additionally, we should integrate out random effects θ_i, $i=1,\ldots,n$. The resulting integral may be cumbersome and hence we also *complete* the likelihood, by conditioning on the random effects. The complete conditional log-likelihood is defined when $\boldsymbol{C}'\boldsymbol{\beta}=0$ and is given by the following expression:

$$\begin{aligned} l_c(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\alpha}) = & \sum_{i=1}^{n}\sum_{j=1}^{S} Y_{ij}\log(p_j(Y_{i,j-1},\ldots,Y_{i1},\theta_i,\boldsymbol{X}_i)) \\ & + (1-Y_{ij})\log(1-p_j(Y_{i,j-1},\ldots,Y_{i1},\theta_i,\boldsymbol{X}_i)) \\ & - \sum_{i=1}^{n}\log(1-\prod_{j=1}^{S}(1-p_j(0,\ldots,0,\theta_i,\boldsymbol{X}_i))) \\ & + \sum_{i=1}^{n}\log(f(\theta_i,\boldsymbol{\alpha})), \end{aligned} \tag{4}$$

with the convention that $p_1(\cdot,\theta_i,\boldsymbol{X}_i)=p_1(\theta_i,\boldsymbol{X}_i)$. For ease of notation we illustrate assuming a continuous distribution is used for θ, where f denotes the density. Latent class models can be accommodated with few adjustments, which are given at the end of the section (see also Coull and Agresti (1999)).

After plug-in of (3) and straightforward algebra, we have:

$$\begin{aligned} l_c(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) &= \sum_{i=1}^{n}\sum_{j=1}^{S} Y_{ij}(\beta_{jY_{i1},\ldots,Y_{i,j-1}} + \theta_i + \boldsymbol{\gamma}'\boldsymbol{X}_i) \\ &- \log(1 + \exp(\beta_{jY_{i1},\ldots,Y_{i,j-1}} + \theta_i + \boldsymbol{\gamma}'\boldsymbol{X}_i)) \\ &- \sum_{i=1}^{n} \log\left(1 - \prod_{j=1}^{S}(1 + \exp(\beta_{j0,\ldots,0} + \theta_i + \boldsymbol{\gamma}'\boldsymbol{X}_i))^{-1}\right) \\ &+ \sum_{i=1}^{n} \log(f(\theta_i|\boldsymbol{\alpha})), \end{aligned} \tag{5}$$

where $\beta_{10,\ldots,0} = \beta_1$.

It is now a matter of setting up a constrained EM algorithm to obtain the MLE. At the E step we must obtain the expected value of (4) with respect to the current posterior distribution of the random effects

$$f(\theta_i|\boldsymbol{\alpha}, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \propto f(\theta_i|\boldsymbol{\alpha}) p(\boldsymbol{Y}|\boldsymbol{X}, \theta_i, \boldsymbol{\beta}, \boldsymbol{\gamma})/(1 - p(\boldsymbol{0}|\boldsymbol{X}, \theta_i, \boldsymbol{\beta}, \boldsymbol{\gamma})), \tag{6}$$

where (6) is evaluated conditionally on the current parameter estimates. The resulting integral is of the form

$$\begin{aligned} Q(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\alpha}) &= \sum_{i=1}^{n} \int \Bigg[\sum_{j=1}^{S} Y_{ij}(\beta_{jY_{i1},\ldots,Y_{i,j-1}} + \theta_i + \boldsymbol{\gamma}'\boldsymbol{X}_i) \\ &- \log(1 + \exp(\beta_{jY_{i1},\ldots,Y_{i,j-1}} + \theta_i + \boldsymbol{\gamma}'\boldsymbol{X}_i)) \\ &- \log\left(1 - \prod_{j=1}^{S}(1 + \exp(\beta_{j0,\ldots,0} + \theta_i + \boldsymbol{\gamma}'\boldsymbol{X}_i))^{-1}\right) \\ &+ \log(f(\theta_i|\boldsymbol{\alpha}))\Bigg] f(\theta_i|\boldsymbol{\alpha}, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}', \boldsymbol{\gamma}')\, d\theta_i. \end{aligned} \tag{7}$$

Given that θ_i is one-dimensional, an accurate and computationally efficient way of computing (7), and similarly the normalizing constant in (6), is given by use of quadrature. A Gauss-Hermite quadrature can be used for Gaussian random effects, while a Gauss–Laguerre quadrature may be better under other assumptions.

At the M step we maximize (7) under the constraint $\boldsymbol{C}'\boldsymbol{\beta} = 0$. To this end, we implement an instance of the Aitchison and Silvey (1958) (AS) algorithm, see also Bergsma (1997) and Lang (2004). The AS algorithm is an iterative quadratic programming algorithm based on indefinite Lagrange multipliers. Optimization of (7) 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 $\boldsymbol{\eta} = (\boldsymbol{\beta}\ \boldsymbol{\gamma})$, $\boldsymbol{s}(\boldsymbol{\eta}|\boldsymbol{\eta}')$ denotes the gradient of (7) with respect to $\boldsymbol{\eta}$, and $\boldsymbol{\lambda}$ is a vector of Lagrange multipliers. 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 Web 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 $\boldsymbol{C}$ with blocks of zeros to obtain $\boldsymbol{D}$, so that the constraint $\boldsymbol{C}'\boldsymbol{\beta} = 0$ is equivalent to $\boldsymbol{D}'\boldsymbol{\eta} = 0$. Suppose $\boldsymbol{D}'\boldsymbol{\eta}_t = 0$. The approximated system of non-linear equations is exactly solved by the updating rule

$$\begin{aligned} \boldsymbol{\eta}_{t+1} = \boldsymbol{\eta}_t &+ \boldsymbol{I}(\boldsymbol{\eta}_t|\boldsymbol{\eta}')^{-1}\boldsymbol{s}(\boldsymbol{\eta}_t|\boldsymbol{\eta}') \\ &- \boldsymbol{I}(\boldsymbol{\eta}_t|\boldsymbol{\eta}')^{-1}\boldsymbol{D}[\boldsymbol{D}'\boldsymbol{I}(\boldsymbol{\eta}_t\boldsymbol{\eta}')^{-1}\boldsymbol{D}]^{-1}[\boldsymbol{D}'\boldsymbol{I}(\boldsymbol{\eta}_t|\boldsymbol{\eta}')^{-1}\boldsymbol{s}(\boldsymbol{\eta}_t|\boldsymbol{\eta}')]. \end{aligned} \tag{8}$$

The AS algorithm might become unstable when the sample size is low or when the number of constraints is close to the number of parameters. In those cases one may include a step-length adjustment, or simply proceed with a numerical M step. See also Evans and Forcina (2013) on this point.

The M step is terminated by an update of $\boldsymbol{\alpha}$ obtained through the maximization of

$$\int \log[f(\theta_i|\boldsymbol{\alpha})] f(\theta_i|\boldsymbol{\alpha}, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}', \boldsymbol{\gamma}')\, d\theta_i.$$

Under the assumption of Gaussian random effects with variance α^2, this is accomplished by the updating rule

$$\alpha^2 = \sum_{i=1}^{n} \frac{\int \theta_i^2 f(\theta_i|\boldsymbol{\alpha}, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}', \boldsymbol{\gamma}')\, d\theta_i}{n},$$

where the integral is approximated using the usual Gauss–Hermite quadrature rule. Similar updating rules are obtained under T and Laplace assumptions, while under logit-Beta assumptions one can use the method of moments. With latent class models, $F(\boldsymbol{\alpha})$ is expressed as $\Pr(\theta_i = \xi_j) = \alpha_j$ for $j = 1, \ldots, k$ and some $k \in \mathcal{N}$ (which can be chosen using AIC). One should therefore estimate both latent locations $\boldsymbol{\xi}$ and the latent masses $\boldsymbol{\alpha}$. The first shall be updated within the AS algorithm, with the same expression as (8) after augmenting $\boldsymbol{s}(\boldsymbol{\eta}|\boldsymbol{\eta}')$ and $\boldsymbol{I}(\boldsymbol{\eta}|\boldsymbol{\eta}')$ with the opportune derivatives, which are given once again in Web Appendix. The latent masses are instead obtained as in (6), based on the updates for the other parameters. Standard errors of the parameters may be obtained through a resampling procedure.

3.2. *Population Size Estimation*

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

$$\hat{N} = \sum_{i=1}^{n} \left(1 - \prod_{j=1}^{S} (1 - p_j(0, \ldots, 0, \hat{\theta}_i, \boldsymbol{X}_i)) \right)^{-1}, \qquad (9)$$

where $p_j(\cdot)$ is obtained after plug-in the MLE for all parameters and $\hat{\theta}_i$ is the empirical Bayes estimate of θ_i, that is,

$$\hat{\theta}_i = \int \theta_i f(\theta_i | \hat{\boldsymbol{\alpha}}, \boldsymbol{Y}, \boldsymbol{X}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) \, d\theta_i$$

where the posterior for θ_i is evaluated at the MLE. The integrals in (9) can be evaluated through the same quadrature rule used at the E step when F is continuous.

In order to obtain the standard error of $\hat{N}$, we let as in van der Heijden et al. (2003) and Böhning (2008) $\delta_i = I(\sum_j Y_{ij} > 0)$ and express (9) as

$$\hat{N} = \sum_{i=1}^{N} \delta_i \left(1 - \prod_{j=1}^{S} (1 - p_j(0, \ldots, 0, \hat{\theta}_i, \boldsymbol{X}_i)) \right)^{-1}.$$

By conditioning,

$$\mathrm{Var}(\hat{N}) = \mathrm{Var}_n(E(\hat{N}|n)) + E_n[\mathrm{Var}(\hat{N}|n)]. \qquad (10)$$

The first term on the right-hand side can be estimated by $\sum_{i=1}^{n} (1 - w_i)/w_i^2$, where $w_i = 1 - \prod_{j=1}^{S} (1 - p_j(0, \ldots, 0, \hat{\theta}_i, \boldsymbol{X}_i))$. The second term is equivalent to the variance of the estimates of N obtained after resampling.

4. Model Selection

When estimation of the population size is not the only issue (see concluding remarks), one may want to proceed as in Farcomeni (2011) and specify a set of interpretable models through equality constraints. In this way, not only an estimate $\hat{N}$ is obtained, but also an explanation of animal behavior. Opportune choices for $\boldsymbol{C}$ give back the classical M_{tbh}, and the use of covariates generalizes it to an M_{hotb}, and all possible submodels. It is also straightforward to obtain models based on homogeneous and inhomogeneous Markov chains of arbitrary order for the conditional capture probabilities (Yang and Chao, 2005). Other possibilities include models based on the *number* of previous capture occasions, possibly with a finite memory, l-th order Markov chains with $l > 1$, delayed onset models (Farcomeni and Scacciatelli, 2013). A geometric progression for behavioral effects is obtained by assuming equality of capture probabilities conditional on the same number of occasions since the last capture. The best model in the set might be selected using an information criterion, like the AIC.

On the other hand, in many cases an estimate of the population size is the only target of the analysis. In that case one may want to explore the model class more broadly, and choose the best fitting model even if its interpretation is not easy. It shall be noted that even for a fixed number of parameters the number of possible constraint matrices is huge, therefore stepwise methods are also not feasible. We propose to proceed along the lines of a LASSO-type penalization, similar to that used by Bondell and Reich (2009) in a completely different framework. More precisely, we substitute the constraint $\boldsymbol{C}'\boldsymbol{\beta} = 0$ with the constraint

$$\sum_{j \le h} \sum_{\boldsymbol{a}} \sum_{\boldsymbol{b}} w_{jh}(\boldsymbol{a}, \boldsymbol{b}) |\beta_{ja_1,\ldots,a_{j-1}} - \beta_{hb_1,\ldots,b_{h-1}}| \le \lambda, \qquad (11)$$

for some $\lambda \ge 0$. The function $w_{jh}(\boldsymbol{a}, \boldsymbol{b})$ is a weight function which can be chosen according to different criteria. One strategy suggested in Bondell and Reich (2009) is to give more weight to differences that are based on a larger number of observations, by fixing

$$w_{jh}(\boldsymbol{a}, \boldsymbol{b}) = \frac{\sqrt{\sum_i I(Y_{i1} = a_1, \ldots, Y_{i,j-1} = a_{j-1}) + \sum_i I(Y_{i1} = b_1, \ldots, Y_{i,h-1} = b_{h-1})}}{2}, \qquad (12)$$

with $w_{jh} = 0$ whenever one of the two sums above is zero. In words, we require that the weighted L_1 norm of the differences between the intercept parameters collected in the vector $\boldsymbol{\beta}$ is bounded by a fixed constant chosen in advance. When $\lambda = 0$, we force $\boldsymbol{\beta}$ to be constant therefore having no behavioral or time effects (M_{ho} model), while as λ is increased more and more parameters are used, and less and less are forced to be equal or at least very close. This is a natural property of the L_1 penalty (Tibshirani, 1996).

Two issues remain: first, how to fit the model (3) under constraint (11). Secondly, how to choose the best λ. The first issue must be tackled by modifying the M step proposed in the previous section. We replace the AS algorithm for updating $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ with a block descent linear programming method. First, $\boldsymbol{\gamma}$ is updated through a classical Fisher-scoring iteration, where the score is derived in Web Appendix in closed form and its gradient is obtained numerically. Then, $\boldsymbol{\beta}$ is updated through the Augmented Lagrangian Adaptive Barrier Minimization Algorithm, which is a numerical method for optimization of smooth nonlinear objective functions with nonlinear constraints. More details can be found in Madsen et al. (2004) and references therein. The remaining part of the M step, and the E step, are analogous to those described in the previous section. As far as a choice for λ is concerned, one can simply repeatedly fit the model for different values of this tuning parameter. In order to choose the best fitting model, we proceed as in Michelot et al. (2013) and use generalized cross-validation. For each possible value of λ, we repeatedly split data in a training (e.g., 90% of the sample) and test

Table 1
RMSE of different models estimated on simulated data. mod = 1 refers to a single and mod = 2 to a double Markov chain model. The estimators are: H: Huggins model. C: Chao, T: Turing estimator. GC: Generalized Chao; GZ: Generalized Zelterman. Results are obtained over B = 1000 replicates.

α	γ	*mod*	S	M_{th}	H	C	T	GC	GZ	$\boldsymbol{C}$	$\boldsymbol{C}^*$
					$N = 50$						
0.25	0.25	1	4	5.17	5.19	7.51	10.65	8.63	1.98	0.46	0.58
0.25	0.25	2	4	2.64	3.27	4.13	6.99	5.25	0.92	0.68	0.78
0.25	0.25	1	6	2.47	1.60	2.89	3.54	3.58	0.14	0.03	0.02
0.25	0.25	2	6	1.05	1.01	1.29	2.08	1.79	0.03	0.11	0.12
0.25	−1.00	1	4	8.20	2.59	7.17	8.27	12.89	8.12	1.31	1.57
0.25	−1.00	2	4	4.68	1.59	4.04	5.07	8.59	5.41	1.78	2.08
0.25	−1.00	1	6	5.28	0.92	4.25	3.28	7.21	2.94	0.78	0.78
0.25	−1.00	2	6	2.56	0.89	2.13	1.86	4.90	2.18	0.83	0.96
0.50	0.25	1	4	5.55	4.54	7.43	9.96	8.57	1.64	0.55	0.61
0.50	0.25	2	4	2.98	2.90	4.45	7.17	5.50	0.97	0.70	0.76
0.50	0.25	1	6	3.04	1.36	3.12	3.46	4.15	0.17	0.05	0.04
0.50	0.25	2	6	1.51	0.96	1.63	2.26	2.16	0.06	0.14	0.16
0.50	−1.00	1	4	9.05	2.13	7.14	7.82	12.62	7.08	1.39	1.51
0.50	−1.00	2	4	4.66	1.32	4.12	4.79	8.18	4.55	1.85	1.95
0.50	−1.00	1	6	5.74	0.84	4.50	3.15	6.63	2.32	0.74	0.73
0.50	−1.00	2	6	3.05	0.96	2.29	1.79	4.37	1.88	0.90	1.04
					$N = 250$						
0.25	0.25	1	4	27.69	21.16	38.01	54.09	40.82	8.21	1.42	1.74
0.25	0.25	2	4	13.57	15.69	21.55	35.96	23.36	2.20	2.33	2.56
0.25	0.25	1	6	13.17	10.06	15.03	17.91	15.88	0.44	0.50	0.52
0.25	0.25	2	6	5.78	4.32	7.01	11.02	7.78	0.31	0.84	0.89
0.25	−1.00	1	4	48.58	9.95	36.81	41.30	60.31	41.41	5.28	7.21
0.25	−1.00	2	4	23.37	3.46	19.07	24.81	41.18	26.06	6.52	7.72
0.25	−1.00	1	6	27.01	1.42	20.05	16.51	29.38	14.08	2.64	3.28
0.25	−1.00	2	6	13.72	1.91	10.61	8.99	18.96	9.68	3.61	4.57
0.50	0.25	1	4	31.15	18.20	37.09	50.43	39.04	5.87	2.46	2.71
0.50	0.25	2	4	15.09	13.10	21.01	33.61	22.75	2.66	3.32	3.79
0.50	0.25	1	6	16.88	8.29	16.92	18.36	17.82	0.70	0.96	0.99
0.50	0.25	2	6	7.71	3.79	8.48	11.41	9.15	0.96	1.27	1.32
0.50	−1.00	1	4	49.71	8.13	34.41	38.11	54.21	31.60	5.59	6.70
0.50	−1.00	2	4	25.11	2.70	18.71	22.85	35.99	18.69	6.69	7.91
0.50	−1.00	1	6	30.05	1.50	21.04	15.81	28.47	11.33	2.95	3.33
0.50	−1.00	2	6	14.85	3.33	9.96	8.06	17.16	5.69	4.01	4.40

set (e.g., the remaining 10%), and evaluate the log-likelihood of the test set at the MLE for the training set. The average over all iterations can be used as a proper scoring rule (Gneiting and Raftery, 2007), and the best λ is set as the maximizer of the score. We have found a good strategy is to start from a very small value, proceed in increments of $\min_{\boldsymbol{a}} \sum_{j \le h} \sum_{\boldsymbol{b}} w_{jh}(\boldsymbol{a}, \boldsymbol{b})$, and stop as soon as the scoring rule decreases for a new candidate penalty value.

5. Simulations

In this section we outline a brief simulation study. We fix the true population size at $N = \{50, 250\}$, Gaussian random effects with $\alpha = \{0.25, 0.5\}$, and also generate one standard normal covariate with $\gamma = \{0.25, 1\}$. We use two options for the M_{bt} part of the true model. In the first case, we specify a first-order Markov chain as in Yang and Chao (2005). The two adjusted log-odds are 1 and -1, leading to trap-shy animals. In the second case, we specify two first-order Markov chains, one for the first $S/2$ and the other for the remaining occasions. For the second Markov chain, we make animals forget their capture history before the $S/2$-th occasion, and specify adjusted log-odds -0.5 and 0.5; leading to slightly trap-happy animals. The number of capture occasions is set at $S = \{4, 6\}$. Overall we then evaluate 32 scenarios. The extent of missingness ($(N - n)/N$) ranges approximately from 1 to 20%.

For each generated data set we fit (3) with known $\boldsymbol{C}$. We also assume, we do not know the true $\boldsymbol{C}$ and use (11) to choose the best λ. The result is indicated by $\boldsymbol{C}^*$. We compare with a classical M_{th} model, which is a special case of (3), using Darroch et al. (1993) mixing distribution; and with Huggins (1989) model (which is an M_{otb} model). Additionally, we fit other options which do not belong to our proposed model class. Specifically, we compare with classical and with covariate-modulated Chao and Zelterman estimators, which should provide an efficient lower bound for the population size; and with Turing estimator. For these estimators refer

Table 2

Coverage of 95% confidence intervals estimated on simulated data. mod = 1 refers to a single and mod = 2 to a double Markov chain model. Results are obtained over B = 1000 replicates.

			$N = 50$		$N = 250$		
α	γ	*mod*	S	$\boldsymbol{C}$	$\boldsymbol{C}^*$	$\boldsymbol{C}$	$\boldsymbol{C}^*$
0.25	0.25	1	4	0.98	0.96	0.99	0.97
0.25	0.25	2	4	0.98	0.98	0.99	0.99
0.25	0.25	1	6	0.96	0.96	0.98	0.98
0.25	0.25	2	6	0.98	0.99	0.97	0.98
0.25	−1.00	1	4	0.97	0.95	0.98	0.97
0.25	−1.00	2	4	0.97	0.96	0.98	0.98
0.25	−1.00	1	6	0.98	0.96	0.99	0.98
0.25	−1.00	2	6	0.98	0.98	0.99	0.97
0.50	0.25	1	4	0.96	0.98	0.93	0.92
0.50	0.25	2	4	0.97	0.98	0.93	0.91
0.50	0.25	1	6	0.90	0.89	0.97	0.92
0.50	0.25	2	6	0.98	0.97	0.92	0.93
0.50	−1.00	1	4	0.96	0.96	0.94	0.95
0.50	−1.00	2	4	0.97	0.98	0.94	0.94
0.50	−1.00	1	6	0.95	0.94	0.94	0.94
0.50	−1.00	2	6	0.94	0.95	0.93	0.94

to Böhning and van der Heijden (2009) and Böhning et al. (2013). Classical M_b and M_{bh} models are not reported after seeing a tendency to break down of the estimates (see Alunni Fegatelli and Tardella (2013) and references therein on this issue). For each model we report the Root Mean Squared Error (RMSE) in estimation of the true population size. The RMSE is estimated as the median over $B = 1000$ replicates. RMSE ($\sqrt{E[(N - \hat{N})^2]}$) of the estimates when $N = 50$ and $N = 250$ are given in Table 1. We also evaluate coverage of the 95% confidence intervals in Table 2.

It can be seen that all competitors considered seem to perform much worse than the proposed estimator both when $\boldsymbol{C}$ is known and when $\boldsymbol{C}$ is estimated using (11). The only exception is GZ, the Zelterman estimator generalized to take into account the covariate, but only in few cases. Estimates based on an unknown $\boldsymbol{C}$ (indicated as $\boldsymbol{C}^*$) have an only slightly larger RMSE than those based on known $\boldsymbol{C}$ and the RMSE seems to decrease at an only slightly slower rate. Coverage is very good in almost all cases.

In order to evaluate the resistance of the estimator to misspecification of the mixing distribution, we still estimate a model based on Gaussian random effects but generate them according to a scaled Student's T distribution with 3 degrees of freedom. The results of this simulation setting are reported in Table 3. Even in this case our proposal achieves very often the lowest RMSE. It can be seen that there is a moderate sensitivity to misspecification, with slightly larger errors. This is well known in the recapture setting (Link, 2003; Farcomeni and Tardella, 2012). Huggins model seems to be somehow more resistant to misspecification, but we speculate it only depends on the specific true mixing distribution chosen. Our model is based on precise parametric assumptions and therefore might sometimes be more sensitive than other ones.

6. Applications

We now illustrate the proposed approach through two real data examples.

6.1. *HIV Data*

The first example we develop is based on the prevalence of HIV-1 infected in the Lazio region of Italy in 1990, based on $S = 4$ lists from four separate testing centers in Rome. Data were first proposed by Abeni et al. (1994), who noted absence of unobserved heterogeneity and significant interaction between the first two lists. The final estimate using a log-linear model with Poisson distribution was $\hat{N} = 12,319$, with 95% CI (9988–14,649). Our approach is particularly amenable to such a dataset, as any kind of list interaction can be represented by constrained conditional capture probabilities.

When comparing models with and without the logit-normal random effect and unknown $\boldsymbol{C}$ we confirm that unobserved heterogeneity should not be included. The best model has $\hat{N} = 11,501$ (95% CI: 9356–13,344), which is slightly lower than the population size estimated by Abeni et al. (1994). The dependency structure with best fit does seem to be slightly more complex than a simple single interaction between two lists. In order to better explore the dependency structure, we work with few options with fixed $\boldsymbol{C}$. To (approximately) reproduce the results in Abeni et al. (1994) we constrain $\beta_{3a} = \beta_{3a'}$, $\beta_{4a''} = \beta_{4a'''}$ for any binary a, a', a'' and a''' of opportune length, but let $\beta_{20} \neq \beta_{21}$. This model is named M_{t12} in Table 4, where we also report on other possible models. Given that $\hat{\beta}_{20}$ and $\hat{\beta}_{3a}$ are rather close for M_{t12}, in Table 4 we also estimate M_{t12c} which further assumes $\beta_{21} = \beta_{311}$. The latter is anyway rejected. Finally, we specify a model M_{t5} which is based on 5 parameters as M_{t12} and on the following assumptions: $\beta_1 = \beta_{310} = \beta_{4001} = \beta_{4011} = \beta_{4111}$, $\beta_{300} = \beta_{301} = \beta_{311} = \beta_{4101}$ and $\beta_{4000} = \beta_{4100} = \beta_{4010} = \beta_{4110}$. The list interaction implied, as anticipated, is rather complex.

Our final estimate with fixed $\boldsymbol{C}$ is $\hat{N} = 11,564$, still slightly smaller than the original estimate of Abeni et al. (1994) but well in agreement with the unknown $\boldsymbol{C}$ situation. We conclude that Abeni et al. (1994) might have slightly over estimated the population size of HIV-1 infected.

6.2. *Snowshoe Hares Data*

The second example we revisit is the snowshoe hares data, where a sample of $n = 68$ hares was observed at least once in $S = 6$ consecutive days. For these data, Coull and Agresti (1999) found support for a normal random effect at logit scale, but observed a strong sensitivity to the dependency structure, with estimates ranging from $\hat{N} = 70$ to $\hat{N} = 92$ for a small choice of four alternative models, with and without heterogeneity. The same data is revisited by Farcomeni and Tardella (2010) through simple M_{th} and M_h models and a semiparametric structure for the random effects. A similar sensitivity to the model structure is found, with estimates ranging from $\hat{N} = 76$ up to $\hat{N} = 89$. Such sensitivity is alarming given that only a very limited number of models is explored in both works. Our approach allows to somehow explore a very large model class. In doing so, we once again fit models with and without unobserved heterogeneity, and for each group we select the best fitting λ. Using generalized cross-validation and the likelihood as scoring

Table 3

RMSE of different models estimated on simulated data when the random effects are misspecified. mod = 1 refers to a single and mod = 2 to a double Markov chain model. The estimators are: H: Huggins model. C: Chao, T: Turing estimator. GC: Generalized Chao; GZ: Generalized Zelterman. Results are obtained over B = 1000 replicates.

					$N = 50$						
α	γ	*mod*	S	M_{th}	H	C	T	GC	GZ	$\boldsymbol{C}$	$\boldsymbol{C}^*$
0.25	0.25	1	4	9.69	1.27	5.86	6.03	7.72	1.99	2.77	2.53
0.25	0.25	2	4	3.55	1.72	2.84	2.88	2.99	2.99	3.24	3.25
0.25	0.25	1	6	5.90	0.82	3.04	2.22	4.06	1.23	1.89	1.92
0.25	0.25	2	6	2.34	1.81	2.26	1.45	2.57	2.00	1.99	2.00
0.25	−1.00	1	4	9.31	2.37	7.11	8.28	13.17	8.84	1.40	1.62
0.25	−1.00	2	4	4.66	1.39	4.50	5.28	10.43	7.06	1.67	1.95
0.25	−1.00	1	6	5.25	0.90	3.68	3.15	6.61	2.79	0.67	0.78
0.25	−1.00	2	6	2.60	0.68	2.20	1.91	4.32	2.04	0.89	1.02
0.50	0.25	1	4	9.53	1.53	5.57	5.74	6.74	2.12	2.79	2.76
0.50	0.25	2	4	4.62	2.18	3.14	3.04	4.09	3.07	3.48	3.47
0.50	0.25	1	6	5.63	1.13	3.50	2.05	4.54	1.68	1.91	1.95
0.50	0.25	2	6	2.86	1.56	2.18	1.40	2.78	2.00	1.97	1.98
0.50	−1.00	1	4	9.38	2.10	7.49	7.88	12.80	6.92	1.60	1.62
0.50	−1.00	2	4	4.47	1.27	3.54	4.53	7.77	4.07	1.69	2.01
0.50	−1.00	1	6	4.99	0.86	3.57	2.96	6.40	2.31	0.76	0.87
0.50	−1.00	2	6	2.47	0.88	2.12	1.71	4.29	2.09	0.96	1.06
					$N = 250$						
0.25	0.25	1	4	51.10	2.84	28.81	27.94	29.64	10.04	15.07	15.36
0.25	0.25	2	4	25.53	7.37	12.71	15.28	15.11	16.02	17.28	17.84
0.25	0.25	1	6	30.33	6.83	17.84	9.50	18.51	9.72	9.91	9.98
0.25	0.25	2	6	11.19	9.95	4.87	2.85	5.34	13.28	12.74	12.89
0.25	−1.00	1	4	48.44	9.70	36.09	41.03	59.50	41.38	3.15	3.99
0.25	−1.00	2	4	23.08	3.10	19.43	24.91	41.38	27.75	4.00	4.58
0.25	−1.00	1	6	26.76	1.80	20.57	16.21	29.29	14.30	1.83	2.02
0.25	−1.00	2	6	14.08	2.07	10.37	9.23	18.66	9.44	2.22	3.06
0.50	0.25	1	4	52.14	3.20	28.49	29.40	30.04	8.85	14.59	14.91
0.50	0.25	2	4	24.11	6.18	11.59	14.32	13.78	16.77	17.27	17.74
0.50	0.25	1	6	29.77	5.96	15.59	8.93	16.60	9.63	9.88	10.00
0.50	0.25	2	6	13.59	8.94	6.03	3.33	6.85	12.58	12.48	12.22
0.50	−1.00	1	4	49.11	8.41	35.13	38.89	55.75	35.12	3.45	4.21
0.50	−1.00	2	4	23.77	3.22	18.84	23.22	38.77	19.08	4.72	5.60
0.50	−1.00	1	6	29.12	1.50	21.01	15.69	29.20	11.97	1.90	2.35
0.50	−1.00	2	6	14.25	2.95	9.67	7.92	18.12	6.23	2.69	3.03

rule, we include the heterogeneity part. The final model structure estimated is rather complex, and difficult to interpretate. On the other hand, we can propose a final estimate $\hat{N} = 77$, associated with a 95% confidence interval of (71–84); which is obtained after exploring a very large model class.

7. Conclusions

We have proposed a completely general M_{hotb} model, based on a conditional parameterization of the capture histories. Working with the conditional rather than the marginal probabilities is a convenient way of dealing with the complex dependency structures that can arise in capture–recapture. Two approaches have been proposed to explore the model class. One possibility is to compare a list of possible models within the large class. The models can be specified through a constraint matrix $\boldsymbol{C}$, or equivalently through a partition of the parameter space (Farcomeni, 2011). If covariates are measured, they can be simply included within the logistic parameterization given that we work with the conditional likelihood. A second possibility we have explored is to maximize the likelihood after penalization of the L_1 distances among the conditional intercepts.

Table 4

HIV data: maximum log-likelihood, AIC, population size estimate with 95% Confidence Interval for different models. M_{t12} includes an interaction between the first two lists, M_{t12c} also assumes that $\beta_{21} = \beta_{311}$.

Model	log-lik	AIC	$\hat{N}$	95% CI
M_0	−3155.92	6313.84	11,529	(9421-13,745)
M_t	−3013.52	6035.04	11,128	(9237-13,019)
M_{t12}	−3010.82	6031.63	12,313	(9787-14,839)
M_{t12c}	−3013.13	6034.26	12,156	(9746-14,566)
M_{t5}	−3009.15	6028.32	11,564	(9234-13,894)

Estimates are derived with an efficient EM algorithm after approximation of the score of the expected complete log-likelihood through Gaussian quadrature. As with any EM algorithm, we recommend a multistart strategy to increase the likelihood of finding the global optimum. We compute standard errors through a resampling strategy, where we initialize the EM only once (from the MLE). When $\boldsymbol{C}$ is unknown, we also keep λ fixed hence ignoring uncertainty about this parameter. This strategy can be seen to work well, as testified for instance by the good coverage of the confidence intervals reported in Section 5. We have found the score in closed form and have made use of numerical differentiation only to compute its first derivative. Numerical first derivatives are usually much more accurate than numerical second derivatives, which we do not need.

In summary, we propose a comprehensive approach to flexible and general capture–recapture models for closed populations, which can deal with any source of heterogeneity and complex time effects and behavioral response. Our approach might be useful not only to estimate a population size, but also to explore the characteristics of the experiment and of the population; based on $\hat{\beta}$ and $\hat{\gamma}$. An experiment in which animals become extremely trap-happy, for instance, may indicate that researchers pet or fed the animals too much. Or, if males are much more likely to be captured than females, we might be able to conclude that there are behavioral differences in the two genders (e.g., males may be more prone to go hunting).

It is worth noting that some formulations of our model, when covariates and/or unobserved heterogeneity are included, are in general over-parameterized and hence not identified. This happens for instance when no constraints are specified for $\boldsymbol{\beta}$. Additionally, as α^2 gets larger and larger the model becomes weakly identified (Coull and Agresti, 1999), as any other M_h model.

There are different venues for further work. One possibility in a Bayesian framework is to explore the use of stochastic search variable selection to obtain the best partition. Another possibility for further work is to let F be unspecified up to minor assumptions (e.g., unimodality). While it is well recognized that a model in which F is not specified is not identifiable (Link, 2003), usually with minor assumptions an estimate of a *lower bound* for N can be obtained. See Holzmann et al. (2006), Mao (2008), Farcomeni and Tardella (2012), and references therein.

8. Supplementary Materials

Web Appendix and code referenced in Sections 1, 3, and 4 are available with this article at the *Biometrics* website on Wiley Online library.

Acknowledgements

The author is grateful to two referees and the editor for kind comments that helped improve the presentation.

References

Abeni, D., Brancato, G., and Perucci, C. A. (1994). Capture-recapture to estimate the size of the population with human immunodeficiency virus type 1 infection. *Epidemiology* **5**, 410–414.

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

Akaike, H. (1973). In *Second International Symposium on Information Theory*, Petrov, B. N. and F., C., (eds), pages 267–281, Budapest: Akademiai Kiado.

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

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

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

Bergsma, W. P. (1997). *Marginal Models for Categorical Data*. Tilburg: Tilburg University Press.

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

Böhning, D. and van der Heijden, P. G. M. (2009). A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations. *Annals of Applied Statistics* **3**, 595–610.

Böhning, D., Vidal-Diez, A., Lerdsuwansri, R., Viwatwongkasem, C., and Arnold, M. (2013). A generalization of Chao's estimator for covariate information. *Biometrics* **69**, 1033–1042.

Bondell, H. D. and Reich, B. J. (2009). Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics* **65**, 169–177.

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

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

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

Darroch, J. N., Fienberg, S. E., Glonek, G. F. V., and Junker, B. W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association* **88,** 1137–1148.

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

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

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

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

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

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102,** 359–378.

Holzmann, H., Munk, A., and Zucchini, W. (2006). On identifiability in capture-recapture models. *Biometrics* **62**, 934–936.

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

Lang, J. (2004). Multinomial Poisson homogeneous models for contingency tables. *Annals of Statistics* **32**, 340–383.

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

Madsen, K., Nielsen, H., and Tingleff, O. (2004). *Optimization With Constraints*. IMM.

Mao, C.-X. (2008). On the nonidentifiability of population sizes. *Biometrics* **64**, 977–979.

Michelot, T., Langrock, R., Kneib, T., and King, R. (2013). Maximum penalized likelihood estimation in semiparametric capture-recapture models. *arXiv:1311.1039* .

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

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

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

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

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society (Series B)* **58,** 267–288.

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

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

Received September 2014. *Revised July* 2015.
Accepted July 2015.