Multiple Systems Estimation for Sparse Capture Data: Inferential Challenges When There Are Nonoverlapping Lists

Lax Chan, Bernard W. Silverman & Kyle Vincent

Taylor & Francis
Taylor & Francis Group

Multiple Systems Estimation for Sparse Capture Data: Inferential Challenges When There Are Nonoverlapping Lists

Lax Chan, Bernard W. Silverman, and Kyle Vincent

Rights Lab, University of Nottingham, Nottingham, UK

ABSTRACT

Multiple systems estimation strategies have recently been applied to quantify hard-to-reach populations, particularly when estimating the number of victims of human trafficking and modern slavery. In such contexts, it is not uncommon to see sparse or even no overlap between some of the lists on which the estimates are based. These create difficulties in model fitting and selection, and we develop inference procedures to address these challenges. The approach is based on Poisson log-linear regression modeling. Issues investigated in detail include taking proper account of data sparsity in the estimation procedure, as well as the existence and identifiability of maximum likelihood estimates. A stepwise method for choosing the most suitable parameters is developed, together with a bootstrap approach to finding confidence intervals for the total population size. We apply the strategy to two empirical datasets of trafficking in US regions, and find that the approach results in stable, reasonable estimates. An accompanying R software implementation has been made publicly available. Supplementary materials for this article are available online.

1. Introduction

Multiple systems estimation, a generalization of the mark-recapture approach (Petersen 1896; Schwarz and Seber 1999), is a class of methods that can be used to estimate the size of hard-to-reach populations in many contexts, including, in recent years, those comprising victims of human trafficking or modern slavery. The methods are typically applied to wildlife populations (Williams, Nichols, and Conroy 2002) and to hidden populations such as injection drug users (King et al. 2013). In the administrative or law enforcement context, multiple systems estimation aims to read across from lists of observed or identified individuals from a study population to estimate the total population of interest (see, e.g., Bales, Hesketh, and Silverman 2015; Cruyff, van Dijk, and van der Heijden 2017). A mathematical model is posited for the pattern of incidences across the lists, and the "dark figure," the number of unobserved cases, is estimated. A survey of the history of the methods and a range of applications is provided, for example, by Bird and King (2018).

Because the method estimates the number of victims including those that are not directly observed or detected, it plays an especially important role in making policy to help combat human trafficking and modern slavery. For example, as set out in Bales, Hesketh, and Silverman (2015), a multiple systems estimate constructed from data collated by a government agency was a key component of the strategy (Home Office 2014) leading to the UK Modern Slavery Act 2015.

A frequent specific challenge posed by data on human trafficking is sparse overlap between the observed administrative lists; indeed, it appears to be the norm rather than the exception that there will be pairs of lists between which there is no observed overlap. This sparsity can lead to inferential and algorithmic difficulties and instabilities if it is not addressed. In applications such as wildlife populations, the researcher may be able to continue capturing from the study population until sufficient overlap is observed between the capture occasions. Such a strategy is not available in the human trafficking context, nor usually in other human rights areas either.

A pair of lists may fail to overlap for a number of reasons: there may be a genuine structural reason why the particular lists cannot overlap; there may be negative correlation between lists; or it may simply be that the overall sample size is relatively small and, especially if the two lists have small capture probabilities, there do not happen to be any cases that are on both lists. In this area, there is as yet limited understanding of data and of mechanisms, and furthermore data are often highly anonymized for reasons of confidentiality and security. Typically, those analyzing the data may not know anything about a list other than an uninformative label, because the collation between lists is carried out by a single trusted individual or agency on that understanding (see, e.g., Bales, Hesketh, and Silverman 2015; Bales, Murphy, and Silverman 2019). Hence, there may be no further information available, beyond simply the number of overlapping cases, as to why no cases are observed in common between two lists.

We approach inference via Poisson log-linear regression modeling applied to counts of individuals that are observed on each possible combination of the lists. This is a well-known technique that allows one to model correlations and

CONTACT Bernard W. Silverman ✉ mail@bernardsilverman.co.uk ▣ Rights Lab, University of Nottingham, Nottingham NG7 2RD, UK.

dependencies between lists. The standard approach was set out by Sandland and Cormack (1984), Cormack (1989), Cormack (1992), Rivest and Daigle (2004), and Bird and King (2018), among others, and implemented in Baillargeon and Rivest (2007) and Rivest and Baillargeon (2019). However, as Fienberg and Rinaldo (2012a) discussed in a much more general context, contingency tables with zero entries, as will arise if there is sparse overlap between lists, may lead to cases where to carry out maximum likelihood estimation it is necessary to extend the range of parameters to include $-\infty$. Even then the maximum likelihood estimate of the model parameters may not exist or may not be identifiable.

In our context, therefore, empty overlaps between lists require careful treatment. The primary objective of this article is to introduce inferential procedures and computational implementations that explicitly handle this case. For simplicity, we focus only on models that include parameters for two-list effects (also called "first-order interactions" by some authors), but the basic concepts of allowing for empty cells, and of checking for the existence of estimates, are straightforward to extend.

We first of all develop a method that fits a model stably, taking proper account of existence and identifiability issues that can arise if the data are sparse. We then consider a model-selection procedure to choose the most suitable set of parameters on which to base inference. A stepwise approach to model selection is used, but so that any effects of choosing the specific model are taken into account in the inference, confidence intervals for the estimation are constructed using the BC_a bootstrap method (DiCiccio and Efron 1996).

The methods are motivated and illustrated by datasets based on human trafficking victims in the New Orleans area (Bales, Murphy, and Silverman 2019) and the Western site of a research study in the USA (Farrell et al. 2019). Simulation studies are used to validate the stepwise approach and are based on datasets generated from these empirical datasets, assigning capture histories to the population members by multinomial sampling, as suggested by Cormack (1992). We conduct our analyses in the R programming language (R Core Team 2016), and have developed an accompanying R software package *SparseMSE* (Chan, Silverman, and Vincent 2019). The package allows readers to implement the methodology on their own data as well as to reproduce the results presented in the article.

The article is organized as follows. Section 2 outlines the Poisson log-linear model and gives the notation and likelihood setup. It details specific issues concerning the existence and identifiability of maximum likelihood estimates, and also discusses issues relating to the breakdown of the assumptions underlying standard likelihood-ratio and information-theoretic approaches. It also develops algorithms for checking efficiently whether models present problems of nonexistence or unidentifiability of estimates. Section 3 develops the model-selection routine and corresponding inference procedure, setting out an efficient algorithmic approach to the bootstrap in this case. A simulation comparison with one of the current standard methods is included. Section 4 presents the results from the two empirical applications, as well as a simulation study informing the choice of threshold in our procedure. The concluding remarks in Section 5 include a comment on the R package,

as well as discussions of the possible extension of the procedure to higher order interactions, and to data with covariate information. There is further detail of several topics in the supplementary materials to the article.

2. The Model

We first define notation and set out the model. The framework leads to an algorithmic approach facilitating correct and stable calculations. We then discuss the implications of sparse counts on existing inferential methods, followed by a discussion on checks for existence of maximum likelihood estimates and identifiability of the model.

2.1. Notation and Definitions

Suppose we have t capture occasions, or lists, on which members of the population can occur. An individual's *capture history* is the set of lists on which the individual is actually observed, or captured. A capture history is a subset $\boldsymbol{\omega}$ of $\{1, 2, \ldots, t\}$.

Now suppose that there are m individuals captured at least once in our study. Denote the m observed capture histories by $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_m$. For any particular capture history $\boldsymbol{\omega}$ define $N_{\boldsymbol{\omega}}$ to be the number of individuals observed to have exactly that capture history, that is, the number of $\boldsymbol{\xi}_i$ equal to $\boldsymbol{\omega}$. It is important to note that the actual data consist of a sample of size m from a discrete distribution over the possible capture histories.

The *order* of a capture history is defined to be the number of captures in the set. The braces are often omitted when the members of the history are given as suffices. Thus, for example, if $t = 4$ the capture history $\{1, 3\}$ has order 2, and $N_{\{1,3\}}$, usually written N_{13}, is the number of individuals that are observed on both lists 1 and 3 but not on lists 2 or 4. A particular capture history of interest, with order 0, is the *null capture history* $\emptyset$. The quantity $N_{\emptyset}$ is the *dark figure* of individuals that are not captured on any list, and therefore cannot be observed. The observed data give rise to the $2^t - 1$ values $\{N_{\boldsymbol{\omega}} : \boldsymbol{\omega} \neq \emptyset\}$, which we will also write as $\mathbf{N}$.

It is characteristic of data collected in the modern slavery context that there will be some capture histories for which the observed count is zero. Typically, each list only records a relatively small proportion of the total population, because of the "hidden" nature of modern slavery as a crime, and the numbers of cases recorded on any particular pattern of overlaps between lists can easily be considerably smaller.

For any capture history $\boldsymbol{\omega}$ define

$$N_{\boldsymbol{\omega}}^* = \sum_{\boldsymbol{\psi} \supseteq \boldsymbol{\omega}, \boldsymbol{\psi} \neq \emptyset} N_{\boldsymbol{\psi}}. \tag{1}$$

Thus, $N_{\boldsymbol{\omega}}^*$ is the number of observed cases that appear on all the lists in $\boldsymbol{\omega}$, regardless of whether they do or do not appear on other lists. For example, N_{12}^* is the total number of cases that are on both lists 1 and 2, while N_{12} is the number of individuals that are on lists 1 and 2 but not on lists 3, 4, …. We will call $\{i, j\}$ a *nonoverlapping pair* of lists if $N_{ij}^* = 0$, so that no individual appears on both lists. The main objective of this article is to develop estimation procedures and algorithms that properly account for these nonoverlapping pairs of lists.

Because of the restriction to $\boldsymbol{\psi} \neq \emptyset$ in the defining sum, the quantity $N^*_\emptyset$ does not include the dark figure but is the sum of all the observed values $N_{\boldsymbol{\psi}}$, the total of individuals actually captured at some point in the study.

2.2. The Poisson Log-Linear Model

A standard model for the analysis is the Poisson log-linear model as set out by Cormack (1989). This assumes that, independently for each $\boldsymbol{\omega}$,

$$N_{\boldsymbol{\omega}} \sim \text{Poisson}(\mu_{\boldsymbol{\omega}}) \quad \text{with} \quad \log \mu_{\boldsymbol{\omega}} = \sum_{\boldsymbol{\theta} \subseteq \boldsymbol{\omega}} \alpha_{\boldsymbol{\theta}} \qquad (2)$$

for certain parameters $\alpha_{\boldsymbol{\theta}}$ indexed by the possible capture histories. Capture histories are used in two different ways, first to index the observed data, and second to index the parameters. Usually, but not invariably, the letter $\boldsymbol{\omega}$ will be used when observations $N_{\boldsymbol{\omega}}$ are indexed and $\boldsymbol{\theta}$ for parameters $\alpha_{\boldsymbol{\theta}}$. The index $\boldsymbol{\psi}$ will be used in either case, as required.

Thus, for example, the dark figure has expected value $\exp \alpha_\emptyset$, while the expected value of N_{13} is $\exp(\alpha_\emptyset + \alpha_1 + \alpha_3 + \alpha_{13})$. Denoting by $\hat{\alpha}_\emptyset$ the maximum likelihood estimate of the parameter $\alpha_\emptyset$, the estimate of the total population size will be $N^*_\emptyset + \exp \hat{\alpha}_\emptyset$, the sum of the total number of cases actually observed and the estimate of the dark figure.

Altogether, there are 2^t parameters $\alpha_{\boldsymbol{\theta}}$, corresponding to the 2^t capture histories including the null capture history. There are only $2^t - 1$ observable data points $N_{\boldsymbol{\omega}}$ from which to estimate the parameters; without placing constraints on the $\alpha_{\boldsymbol{\theta}}$ parameters, the model is not identifiable. As Cormack (1989) set out, the natural approach is to set some of the $\alpha_{\boldsymbol{\theta}}$ to zero, and then to estimate the remainder by maximum likelihood; for example, one may set all coefficients indexed by third- or higher order histories to zero, and we will do this throughout. Even if all the two-list coefficients (those indexed by pairs of lists) are included, the number of parameters to be estimated is $1 + t + \frac{1}{2}t(t-1) \leq 2^t - 1$ provided $t \geq 3$. Model choice then reduces to deciding which two-list coefficients to include, and will be discussed further in Section 3. For any particular choice of coefficients, the estimation can be put into a standard generalized linear model formulation.

A consequence of the definitions (1) and (2) is that, for each $\boldsymbol{\omega}$,

$$N^*_{\boldsymbol{\omega}} \sim \text{Poisson}(\mu^*_{\boldsymbol{\omega}}) \quad \text{where} \quad \mu^*_{\boldsymbol{\omega}} = \sum_{\boldsymbol{\psi} \supseteq \boldsymbol{\omega}, \boldsymbol{\psi} \neq \emptyset} \mu_{\boldsymbol{\psi}}.$$

Unlike the $N_{\boldsymbol{\omega}}$, the $N^*_{\boldsymbol{\omega}}$ are not independent. For example, if capture histories $\boldsymbol{\omega}$ and $\boldsymbol{\psi}$ share any lists, then the variables $N^*_{\boldsymbol{\omega}}$ and $N^*_{\boldsymbol{\psi}}$ will be dependent.

2.3. The Log-Likelihood Function

Before considering the treatment of nonoverlapping pairs of lists, we derive some properties of the likelihood function. Let Θ be the collection of indices of parameters included in the model, and $\boldsymbol{\alpha} = (\alpha_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta)$ the vector of parameters to be estimated. Note that Θ always contains $\emptyset$. Up to an additive

constant depending only on the data, the log-likelihood is given by

$$\ell(\boldsymbol{\alpha}|\mathbf{N}) = \sum_{\boldsymbol{\omega} \neq \emptyset} \{N_{\boldsymbol{\omega}} \log(\mu_{\boldsymbol{\omega}}) - \mu_{\boldsymbol{\omega}}\}.$$

Substitute the definition of the model, reverse the order of summation, and then substitute the definition (1), to obtain

$$\sum_{\boldsymbol{\omega} \neq \emptyset} N_{\boldsymbol{\omega}} \log(\mu_{\boldsymbol{\omega}}) = \sum_{\boldsymbol{\omega} \neq \emptyset} \left\{ N_{\boldsymbol{\omega}} \sum_{\boldsymbol{\theta} \subseteq \boldsymbol{\omega}, \boldsymbol{\theta} \in \Theta} \alpha_{\boldsymbol{\theta}} \right\} = \sum_{\boldsymbol{\theta} \in \Theta} \left\{ \alpha_{\boldsymbol{\theta}} \sum_{\boldsymbol{\omega} \supseteq \boldsymbol{\theta}, \boldsymbol{\omega} \neq \emptyset} N_{\boldsymbol{\omega}} \right\}$$
$$= \sum_{\boldsymbol{\theta} \in \Theta} \alpha_{\boldsymbol{\theta}} N^*_{\boldsymbol{\theta}}. \qquad (3)$$

Turning to the other term in the log-likelihood,

$$-\sum_{\boldsymbol{\omega} \neq \emptyset} \mu_{\boldsymbol{\omega}} = \sum_{\boldsymbol{\omega} \neq \emptyset} \left\{ -\exp\left[\sum_{\boldsymbol{\theta} \subseteq \boldsymbol{\omega}, \boldsymbol{\theta} \in \Theta} \alpha_{\boldsymbol{\theta}} \right] \right\} = C(\boldsymbol{\alpha}),$$

say. Regarded as a function of the $\alpha_{\boldsymbol{\theta}}$, each $\mu_{\boldsymbol{\omega}}$ is an increasing function of each of its arguments, and hence $C(\boldsymbol{\alpha})$ is a decreasing function of each of its arguments $\{\alpha_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. Furthermore, $C(\boldsymbol{\alpha})$ is a sum of concave functions of linear combinations of its arguments, so $\ell(\boldsymbol{\alpha}|\mathbf{N})$ is the sum of a linear and a concave function, and hence is a concave function. However, as Fienberg and Rinaldo (2012a) showed in a much more general and abstract context, and as we shall see below, the maximum likelihood estimate of $\boldsymbol{\alpha}$ need not be unique or even exist at all.

The expressions for the components of the log-likelihood function demonstrate the following, which will be useful in our discussion of model choice:

1. The statistics $\{N^*_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ are jointly sufficient for the parameters $\boldsymbol{\alpha}$.
2. Given any $\boldsymbol{\omega}$ in Θ, $N^*_{\boldsymbol{\omega}}$ is sufficient for $\alpha_{\boldsymbol{\omega}}$ if all the other parameters $\{\alpha_{\boldsymbol{\psi}} : \boldsymbol{\psi} \in \Theta, \boldsymbol{\psi} \neq \boldsymbol{\omega}\}$ are kept fixed.

2.4. Dealing With Nonoverlapping Pairs

Suppose that $\{i, j\}$ is a nonoverlapping pair, so that $N^*_{ij} = 0$, and that α_{ij} is one of the parameters in the model being fitted, so that $\{i, j\} \in \Theta$. In the terminology of Fienberg and Rinaldo (2012a) we allow an extended maximum likelihood estimate, which means that that the parameters may take values in $[-\infty, \infty)$. If a parameter $\alpha_{\boldsymbol{\theta}} = -\infty$ then we will have $\mu_{\boldsymbol{\omega}} = 0$ for all $\boldsymbol{\omega} \supseteq \boldsymbol{\theta}$, so the actual Poisson parameters will still all be finite. This section gives an elementary recapitulation of some of the results (Fienberg and Rinaldo 2012a) cast into our specific framework.

In the first term (3) of the log-likelihood, the coefficient of α_{ij} is zero, so the maximum likelihood estimate of α_{ij} will be obtained by maximizing $C(\boldsymbol{\alpha})$. Because $C(\boldsymbol{\alpha})$ is a decreasing function of each of its arguments, whatever the value of the other parameters the likelihood will be maximized as $\alpha_{ij} \to -\infty$. The maximum likelihood estimate of α_{ij} may therefore be regarded as $\alpha_{ij} = -\infty$. This explains why existing software packages yield errors or warnings if there are nonoverlapping pairs in the data and the corresponding parameters are in the model. Because the linear model is expressed in terms of the logarithm of the

Poisson parameter, the value $-\infty$ for α_{ij} gives the value zero for $\mu_{\boldsymbol{\omega}}$ for all $\boldsymbol{\omega} \supseteq \{i, j\}$, a legitimate value for the actual Poisson parameters, regarding a Poisson distribution with parameter zero to be the degenerate distribution with value zero.

Substituting these zeroes for $\mu_{\boldsymbol{\omega}}$ back into the expression for the log-likelihood yields, writing $\boldsymbol{\alpha}_{ij}^{\dagger}$ for the vector of parameters with α_{ij} excluded,

$$\ell(\boldsymbol{\alpha}_{ij}^{\dagger}|\mathbf{N}, \alpha_{ij} = -\infty) = \sum_{\boldsymbol{\omega} \neq \emptyset, \boldsymbol{\omega} \not\supseteq \{i,j\}} \{N_{\boldsymbol{\omega}} \log(\mu_{\boldsymbol{\omega}}) - \mu_{\boldsymbol{\omega}}\}.$$

This is exactly the Poisson log-likelihood based on all the observations except those for the 2^{t-2} capture histories that include both i and j. Note that the sum is over $\boldsymbol{\omega}$ that do not include the set $\{i, j\}$, in other words both of i and j. If there is more than one nonoverlapping pair in Θ, the same calculations can be carried out for each pair, leading to the following algorithm:

1. Initially define $\Omega^{\dagger}$ be the set of all nonnull capture histories and $\Theta^{\dagger} = \Theta$.
2. For each $\{i, j\}$ in Θ for which $N_{ij}^{*} = 0$, record that the maximum likelihood estimator of α_{ij} is $-\infty$ and remove α_{ij} from the set of parameters $\Theta^{\dagger}$ yet to be estimated.
3. For each such $\{i, j\}$ also remove from $\Omega^{\dagger}$ all $\boldsymbol{\omega}$ for which $\boldsymbol{\omega} \supseteq \{i, j\}$ (because $N_{ij}^{*} = 0$ the corresponding $N_{\boldsymbol{\omega}}$ will all be zero).
4. Use the standard generalized linear model approach to estimate the parameters with indices in $\Theta^{\dagger}$ from the observed counts of the capture histories in $\Omega^{\dagger}$. The set $\Theta^{\dagger}$ comprises all the two-list parameters in the model that are not estimated to be $-\infty$.

In the next section, we will see that the final step should also involve an explicit check for the existence and identifiability of the parameter estimates.

2.5. How Existing Methods Go Wrong

Where there is a pair of nonoverlapping lists, existing methods typically iterate toward a large negative estimate for the corresponding parameter α_{ij}, only stopping because the number of iterations exceeds a prescribed limit, or because the second derivative of the log-likelihood is numerically nearly zero. An error or warning message may be produced. By contrast, our approach deals explicitly with α_{ij}, immediately giving it the value that maximizes the likelihood over the extended range $[-\infty, \infty)$. Once the parameters corresponding to nonoverlapping pairs of lists have been correctly estimated, all the other parameters are estimated by an iterative process that converges rapidly and does not yield any errors.

Suppose, for the moment, that a large negative value of α_{ij} is used, say $\alpha_{ij} = -20$ rather than $\alpha_{ij} = -\infty$. For practical purposes e^{-20} is zero, so the fitted values of $\mu_{\boldsymbol{\omega}}$ will be essentially zero for all $\boldsymbol{\omega} \supseteq \{i, j\}$ and the corresponding terms will make no contribution to the maximization of the likelihood of the other parameters. Hence, the fitted values of the other parameters will be much the same as in our approach, which actually estimates the parameter α_{ij} correctly. We are not fitting a different model than other approaches; rather, we are correctly fitting a model that other approaches can only fit approximately and in an unsatisfactory way.

Table 1. Comparison of the performance of standard approaches using `glm` with the method set out in this article.

Dataset	Pair	Standard			Proposed	
		Estimate	SE	p-value	Estimate	p-value
Netherlands	I:K	−20.79	5778	0.997	−∞	9.1×10^{-4}
	K:R	−19.96	2783	0.994	−∞	2.1×10^{-5}
UK	LA:GP	−19.08	5350	0.997	−∞	0.13
	LA:NCA	−19.19	7968	0.998	−∞	0.30

NOTE: The quantity estimated is the parameter α_{ij} for the nonoverlapping pair (i, j) under consideration. The model fitted includes all two-list parameters.

Not only is it inelegant to use an iterative method to approach a known $-\infty$ value of a parameter, but it leads to misleading estimates of the precision of the parameter estimates. Because the second derivative of the log-likelihood also rapidly tends to zero, the estimated parameter tends to have very large reported standard error, suggesting that its estimate is essentially uninformative. Furthermore, standard packages use approaches to inference and model choice based on likelihood and information criteria. The asymptotic theory and arguments behind these approaches, for example, Wilks (1938) and Akaike (1974), break down when parameters are at an extreme of their ranges, as is the case in our application for the parameters corresponding to nonoverlapping lists (see Section 1 of the supplementary materials for a simulation example illustrating that the likelihood asymptotics do not hold).

An exploration of the possibly misleading precision estimates, for two real datasets, is given in Table 1. The datasets are from the UK (Home Office 2014) and The Netherlands (van Dijk et al. 2017), both tabulated in Silverman (2020). In each case the data consist of six lists, and in both cases there are two nonoverlapping pairs. We will see that the corresponding parameters are significant in one case but not the other. The standard errors and p-values for `glm` are those produced using the default method for that routine.

The table shows the result of fitting the model including all two-list effects, using two algorithms, one being a "standard" approach (Rivest and Baillargeon 2019) that makes use of the R program `glm`, and the other the method set out above. In the standard approach, the call to `glm` actually records convergence, but after 21 and 22 iterations, respectively, which is close to the default maximum number 25 of iterations in `glm`. In both cases a warning is generated. By contrast, the call to `glm` within our approach only requires 6 or 7 iterations. The estimates of all the other parameters, as expected, are virtually the same in both cases. The p-value for our approach is the probability that the nonoverlap of the relevant pair could have occurred by chance when the model is fitted without the corresponding parameter; see Section 3.1. It can be seen that the effects are highly significant for the Netherlands data but not significant for the UK data. The reported standard errors and p-values are not meaningful for the standard approach.

In fact there are additional aspects not handled by current methods that need to be addressed, even if one allows for the parameters to be estimated over the extended range $[-\infty, \infty)$, and these are discussed in Section 2.6.

2.6. Existence and Identifiability Issues

Two estimability issues may arise when applying multiple systems estimation to sparse data, both of which will mean that the model will not give a well-defined finite estimate of the population size.

One possibility is that there is no value of the parameter vector $\boldsymbol{\alpha}$ that maximizes the likelihood, even allowing the extended range $[-\infty, \infty)$ for the parameters. The other, separate, possibility is that (whether or not the likelihood can be maximized) there is parameter redundancy and the estimates are not identifiable. We discuss the existence question first.

Fienberg and Rinaldo (2012b) showed that existence of the estimate can be checked by solving a linear programming problem. Defining $\Theta^\dagger$ and $\Omega^\dagger$ as in the algorithm set out in Section 2.4, let $\mathbf{A}$ be the incidence matrix that maps the parameters in $\Theta^\dagger$ to the logarithm of the expected values of the counts of capture histories in $\Omega^\dagger$. From (2), for $\boldsymbol{\theta} \in \Theta^\dagger$ and $\boldsymbol{\omega} \in \Omega^\dagger$, $\mathbf{A}_{\boldsymbol{\omega\theta}} = 1$ if $\boldsymbol{\theta} \subseteq \boldsymbol{\omega}$ and 0 otherwise. Let $\mathbf{t}$ be the vector of sufficient statistics $N_{\boldsymbol{\theta}}^*$ for $\boldsymbol{\theta} \in \Theta^\dagger$. Then set up the linear programming problem of finding the maximum value of s over all scalars s and all real vectors $\mathbf{x} = (x_{\boldsymbol{\omega}}, \boldsymbol{\omega} \in \Omega^\dagger)$ satisfying the constraints

$$\mathbf{A}^T \mathbf{x} = \mathbf{t} \quad \text{and} \quad x_{\boldsymbol{\omega}} - s \geq 0 \text{ for all } \boldsymbol{\omega} \in \Omega^\dagger. \quad (4)$$

A necessary and sufficient condition for a maximum likelihood estimate of $\boldsymbol{\alpha}$ to exist (possibly allowing some parameters to be $-\infty$) is that the maximizing value $s_{\max}$ of s is strictly greater than 0.

Setting $x_{\boldsymbol{\omega}} = N_{\boldsymbol{\omega}}$ for all $\boldsymbol{\omega}$ and $s = \min N_{\boldsymbol{\omega}}$ will yield a feasible solution satisfying (4). Hence, $s_{\max}$ will be at least the minimum of the observed $N_{\boldsymbol{\omega}}$ over $\Omega^\dagger$. In the nonsparse case, where every combination of capture histories is observed at least once, this minimum will be strictly positive and hence the maximum likelihood estimator will always exist.

The other possibility is that, even if the likelihood can be maximized, the parameters are nonidentifiable, so that the estimate is not unique, a state of affairs also called parameter redundancy (see, e.g., Far, Papathomas, and King 2019). The model will be identifiable if and only if $\mathbf{A}$ is of full column rank. We show in Section 3 of the supplementary materials that nonidentifiability can only arise if all list pairs are in the model and if the data are so sparse that every set of three lists contains at least one nonoverlapping pair. This condition is easily checked.

Fienberg and Rinaldo (2012a) point out that most or all standard generalized linear modeling packages fail to check for existence of estimates. Nor do programs necessarily report unidentifiability directly, more often arbitrarily removing one or more of the parameters. Unless every possible capture history is actually represented in the observed data, therefore, it is important to check that a potential model gives a strictly positive value for the linear programming problem. If the full model containing all two-list parameters is being fitted then, in addition, identifiability should also be checked. If the model fails on either count it should be ruled out. These checks incur only a small computational overhead.

A simple example is given in Table 2. As there are three possible two-way interactions, there are $2^3 = 8$ possible choices of the two-list parameters to include in the model. We summarize the linear programming output $s_{\max}$ and test results in Table 3.

Table 2. An artificial dataset with three lists.

Cases observed only on one list		Cases observed on exactly 2 lists	
List	Number	Lists	Number
A	40	A&B	6
B	30		
C	20		

NOTE: In this dataset, there are no cases with capture histories AC, BC, or ABC.

Table 3. Summary of linear programming output and test result for all possible choices of two-list effects to include in the model.

Two-list parameters included	Test result	$s_{\max}$
None	No error	1.2
α_{AB}	Nonexistent MLE	0
α_{AC}	No error	3
α_{BC}	No error	3
α_{AB}, α_{AC}	Nonexistent MLE	0
α_{AB}, α_{BC}	Nonexistent MLE	0
α_{AC}, α_{BC}	No error	6
$\alpha_{AB}, \alpha_{AC}, \alpha_{BC}$	Unidentifiable	6

NOTE: For the model containing all three two-list effects, there are finite values of the Poisson means $\mu_{\boldsymbol{\omega}}$ that maximize the likelihood, so $s_{\max} > 0$, but these do not correspond to unique values of parameters in the model.

The results show that there is no immediate hierarchical relationship between models that do or do not satisfy the criterion for estimates to exist. For example, the linear program result is zero for the model including AB and AC, but either adding the third effect BC, or removing AB, will yield a model for which the result is strictly positive. This issue is elucidated further in Section 2.7.

2.7. Checking All Models

Given a particular dataset, it is useful in certain contexts to check that the estimates exist no matter which two-list terms are included in the model. An appropriate algorithm allows the Fienberg–Rinaldo conditions to be confirmed much more quickly than the brute force approach of simply checking the criteria for every possible model. It will be assumed throughout that the model contains the intercept parameter $\alpha_\emptyset$ and the main effect parameters α_i for $i = 1, \ldots, t$. The model choice to be made is which, if any, of the two-list parameters α_{ij} also to include. Because there are $\frac{1}{2}t(t-1)$ pairs $\{i, j\}$, the number of possible models is $2^{t(t-1)/2}$, which rapidly becomes very large as the number of lists increases.

Suppose that $\{i, j\}$ is an overlapping pair of lists, in that $N_{ij}^* > 0$, and that the parameter α_{ij} is in the current parameter set Θ. Consider the effect of removing this parameter from the model. Because $\{i, j\}$ is an overlapping pair, this will not change the set $\Omega^\dagger$, but because $\{i, j\}$ is removed from Θ it will also be removed from $\Theta^\dagger$ (again defining $\Theta^\dagger$ and $\Omega^\dagger$ as in Section 2.4). In the linear programming problem (4), this will remove one column from the matrix $\mathbf{A}$ and the corresponding element of $\mathbf{t}$. Hence, one constraint will be removed, and therefore the maximum value of s cannot decrease. Therefore, if the estimate exists for parameter set Θ it will necessarily exist for subsets of Θ obtained by removing overlapping pairs. It follows that, to confirm whether all models satisfy the conditions for estimates to exist, it is only necessary initially to test parameter sets Θ that

include all overlapping pairs, together with a subset (possibly empty) of the nonoverlapping pairs. If there are M nonoverlapping pairs in the data, then the number of such models is 2^M; solving the linear programming problem for all these models is now feasible for a much larger range of datasets than if all $2^{t(t-1)/2}$ models have to be considered explicitly. If the estimates exist for all such models, then they will exist for all models.

These checks were carried out for all the datasets discussed in this article. For the full UK and Netherlands data, the number of models to be checked by solving a linear programming problem is reduced by a factor of 8192. Details for two other datasets are given in Sections 4.1 and 4.2. In every case, in contrast to the example set out in Table 2, the extended maximum likelihood estimate exists and is unique for every possible choice of model.

In the event that there are models for which the estimate does not exist, the approach can be extended to find a list of all such models efficiently. Let Θ_1 be the set of parameter indices corresponding to the empty capture history and all capture histories of order 1, and let Θ_2^{over} and Θ_2^{non} be those corresponding to overlapping and nonoverlapping pairs, respectively. The initial search is over all models containing both Θ_1 and Θ_2^{over}. Suppose it yields a subset $\tilde{\Theta}_2^{\text{non}}$ with the property that there is no maximum likelihood estimate within the model with parameter set $\Theta_1 \cup \Theta_2^{\text{over}} \cup \tilde{\Theta}_2^{\text{non}}$. We then perform a hierarchical search, retaining $\Theta_1 \cup \tilde{\Theta}_2^{\text{non}}$, over models where overlapping pairs are removed. At the first stage, parameters in Θ_2^{over} are removed individually and each resulting model checked. If any such model yields a zero result in the linear program, that is recorded, and the possibility of removing a second overlapping pair is investigated, and so on. At each stage, if the linear program yields a positive result so that the estimate exists, there is no need to investigate that branch of the hierarchy any further.

3. Inference and Model Choice

We now consider how to assess the significance of any particular two-list parameter, and develop a forward stepwise approach to model choice. We also develop the bootstrap procedure for evaluating confidence intervals, and present simulation results comparing the bootstrap with an approach that carries out inference conditional on the model actually selected.

3.1. Calculating Significance

Given any model defined by parameter set Θ, for any $\boldsymbol{\omega}$ define

$$\hat{\mu}_{\boldsymbol{\omega}}[\Theta] = \exp\left(\sum_{\boldsymbol{\theta} \subseteq \boldsymbol{\omega}, \boldsymbol{\theta} \in \Theta} \hat{\alpha}_{\boldsymbol{\theta}}\right),$$

where the $\hat{\alpha}_{\boldsymbol{\theta}}$ are the maximum likelihood estimates of the $\alpha_{\boldsymbol{\theta}}$. Further, define

$$\hat{\mu}_{\boldsymbol{\omega}}^*[\Theta] = \sum_{\boldsymbol{\psi} \supseteq \boldsymbol{\omega}, \boldsymbol{\psi} \neq \emptyset} \hat{\mu}_{\boldsymbol{\psi}}[\Theta].$$

Under these definitions, $\hat{\mu}_{\boldsymbol{\omega}}[\Theta]$ and $\hat{\mu}_{\boldsymbol{\omega}}^*[\Theta]$ are the estimated expected values of $N_{\boldsymbol{\omega}}$ and $N_{\boldsymbol{\omega}}^*$, respectively. The notation $[\Theta]$ makes explicit the dependence on the parameter set Θ.

First, consider how to deal with nonoverlapping pairs within the data. Suppose that for some $\boldsymbol{\theta} \in \Theta$ that $N_{\boldsymbol{\theta}}^* = 0$. Should we actually include $\boldsymbol{\theta}$ in the model? We test the null hypothesis that $\alpha_{\boldsymbol{\theta}} = 0$, which is equivalent to saying that $\boldsymbol{\theta}$ is not in the model. We fit the model without $\boldsymbol{\theta}$ and then consider the p-value of a test statistic. A natural test statistic is $N_{\boldsymbol{\theta}}^*$, because of the results on sufficient statistics in Section 2.3. Recall that this is also a Poisson random variable since it is the sum of independent Poisson random variables (see (1) and (2)). Hence, we test whether 0 is a surprising value to observe for a Poisson distribution estimated from the data but leaving out the parameter indexed by $\boldsymbol{\theta}$. If $\boldsymbol{\theta}$ is in the model, then the observed value has probability one if $\boldsymbol{\theta}$ takes its estimated value.

Hence, proceed as follows:

1. Fit the model leaving out the parameter $\alpha_{\boldsymbol{\theta}}$, in other words using just the parameter set $\Theta \setminus \boldsymbol{\theta}$. For the resulting fitted model, find the estimate $\hat{\mu}_{\boldsymbol{\theta}}^*[\Theta \setminus \boldsymbol{\theta}]$.
2. The estimated parameter has p-value $\exp(-\hat{\mu}_{\boldsymbol{\theta}}^*[\Theta \setminus \boldsymbol{\theta}])$. This is the estimated probability that $N_{\boldsymbol{\theta}}^* = 0$ in the model defined by $\Theta \setminus \boldsymbol{\theta}$.

Unless we have already checked that the parameter set $\Theta \setminus \boldsymbol{\theta}$ passes the linear program test for the existence of the maximum likelihood estimate, that should be done; if the model fails that test then the effective p-value is zero because the parameter $\alpha_{\boldsymbol{\theta}}$ cannot be removed from the model.

This approach can be generalized to construct a (one-sided) p-value for *any* parameter $\boldsymbol{\theta} \in \Theta$ whether or not $N_{\boldsymbol{\theta}}^* = 0$. The p-value is the minimum of $F_{\text{Poiss}}(N_{\boldsymbol{\theta}}^*, \hat{\mu}_{\boldsymbol{\theta}}^*[\Theta \setminus \boldsymbol{\theta}])$ and $\tilde{F}_{\text{Poiss}}(N_{\boldsymbol{\theta}}^*, \hat{\mu}_{\boldsymbol{\theta}}^*[\Theta \setminus \boldsymbol{\theta}])$. Here $F_{\text{Poiss}}(n, \lambda)$ is the lower tail probability that a $\text{Poiss}(\lambda)$ random variable X satisfies $X \leq n$, while $\tilde{F}_{\text{Poiss}}(n, \lambda)$ is the probability that $X \geq n$.

An alternative approach is to use the sufficient statistic $N_{\boldsymbol{\theta}}^*$ for $\alpha_{\boldsymbol{\theta}}$ evaluated against its distribution conditional on the observed values of the sufficient statistics in the model with parameters indexed by $\Theta \setminus \boldsymbol{\theta}$, rather than, as we have, against its unconditional distribution on the estimated model. The conditional distribution does not seem to be easily tractable, but this is an interesting avenue for future research.

3.2. Model Fitting

The model-fitting procedure is detailed stepwise, as follows:

- Step 1: Set a threshold value for the p-value and fit the model with the main effects parameters only.
- Step 2: Consider in turn each two-list parameter not already added to the model, and check that adding it to the model would not lead to a nonexistent estimate (or to nonidentifiability if the full two-way model is proposed).
- Step 3: Among those parameters that pass the checks, find the one with the smallest p-value, using the approach set out in Section 3.1. If that p-value is less than or equal to the given threshold, add the parameter to the model, and go back to Step 2. If the p-value is greater than the threshold, finish.

Note that in Step 2 all two-list parameters not already included are considered, whether the pairs they correspond to are overlapping or nonoverlapping. The method is akin to

forward stepwise regression. Note also that if the algorithm set out in Section 2.7 has already demonstrated that nonexistence and nonidentifiability cannot arise for any model for the dataset in question, then the check in Step 2 is not necessary.

It remains to choose the threshold p-value. We conduct a detailed simulation study in Section 4.3 that points to the choice $p = 0.02$, and that is the value which we would suggest, but users might wish to explore the sensitivity of the result to adjusting the parameter.

3.3. Bootstrapping to Find Confidence Intervals

In general, current approaches find confidence intervals for the population size conditional on the terms actually included in the model, either for the Poisson log-linear model itself, or for modifications such as the multinomial model considered by Sandland and Cormack (1984). Because the choice of model itself depends on the observed data, it is preferable to construct confidence intervals that take account directly of the effect of model selection. A natural way of doing this is to use a bootstrap approach, which will also take account of any biases that the model selection approach may introduce. The BC_a methodology of DiCiccio and Efron (1996) gives second-order accuracy and does not depend on any transformation of the scale on which the data are observed and the estimate of the total population made.

The observed data in our case are the original m observed capture histories $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_m)$. To construct each bootstrap sample, we could draw a random sample $(\boldsymbol{\xi}_1^{\text{boot}}, \boldsymbol{\xi}_2^{\text{boot}}, \ldots, \boldsymbol{\xi}_m^{\text{boot}})$ of size m, with replacement, from the original data. If we denote by $N_{\boldsymbol{\omega}}^{\text{boot}}$ the number of times the capture history $\boldsymbol{\omega}$ occurs in the bootstrap sample, then the $N_{\boldsymbol{\omega}}^{\text{boot}}$ have a multinomial distribution corresponding to m trials and probabilities proportional to the original $N_{\boldsymbol{\omega}}$. In practice, therefore, the $\boldsymbol{\xi}_i^{\text{boot}}$ are not actually constructed, but we sample direct from the multinomial distribution of the capture history totals. The parameter for the number of trials in the multinomial distribution is the number m of capture histories actually observed and does not depend on any estimate of the dark figure.

For each bootstrap sample, we carry out the stepwise fitting procedure and obtain an estimate (bootstrap replication) of the population size. There is no constraint on choosing the same model. The BC_a confidence intervals use percentiles of the bootstrap distribution of the population size, but they adjust the percentile actually used. The adjusted percentiles depend on an estimated bias parameter $\hat{z}_0$, defined so that $\Phi(\hat{z}_0)$ is the proportion of the bootstrap estimates that fall below the estimate from the original data, and an estimated acceleration factor $\hat{a}$, whose derivation depends on a jackknife approach.

The jackknife requires the population size to be estimated from every sample constructed from the original data by leaving out one of the data points $\boldsymbol{\xi}_i$. However, the number of jackknife estimates that need to be evaluated can usually be dramatically reduced, making for considerable computational savings, because the number of distinct values taken by the $\boldsymbol{\xi}_i$, the number of different capture histories actually observed, is in general much smaller than m. If there are K capture histories for which $N_{\boldsymbol{\omega}} > 0$, only K jackknife estimates actually have

to be calculated. These are then weighted in the calculations by the number of times $N_{\boldsymbol{\omega}}$ that the value $\boldsymbol{\omega}$ appears in the original sample. To be explicit, let $\hat{\theta}_{(i)}$ be the estimate of the population size constructed from the original sample leaving out capture history $\boldsymbol{\xi}_i$. The effect of leaving out that capture history is to reduce $N_{\boldsymbol{\xi}_i}$ by one, and so $\hat{\theta}_{(i)} = \hat{\theta}_{\boldsymbol{\xi}_i}^{(-1)}$, where, for each capture history $\boldsymbol{\omega}$ actually observed in the data, $\hat{\theta}_{\boldsymbol{\omega}}^{(-1)}$ is the estimate of the population size from the original sample but with $N_{\boldsymbol{\omega}}$ replaced by $N_{\boldsymbol{\omega}} - 1$. Only the K values $\hat{\theta}_{\boldsymbol{\omega}}^{(-1)}$ have to be calculated.

To calculate the acceleration factor, let $\hat{\theta}_{(\cdot)}$ be the average of the jackknife estimates $\hat{\theta}_{(i)}$. Then

$$\hat{\theta}_{(\cdot)} = m^{-1} \sum_{\boldsymbol{\omega}} \sum_{i : \boldsymbol{\xi}_i = \boldsymbol{\omega}} \hat{\theta}_{(i)} = m^{-1} \sum_{\boldsymbol{\omega} : N_{\boldsymbol{\omega}} > 0} N_{\boldsymbol{\omega}} \hat{\theta}_{\boldsymbol{\omega}}^{(-1)}.$$

Applying a similar weighting argument to the defining equations (6.6) and (6.7) of DiCiccio and Efron (1996), the estimated acceleration factor $\hat{a}$ is then given by

$$\hat{a} = \frac{1}{6} \left\{ \sum_{\boldsymbol{\omega} : N_{\boldsymbol{\omega}} > 0} N_{\boldsymbol{\omega}} (\hat{\theta}_{(\cdot)} - \hat{\theta}_{\boldsymbol{\omega}}^{(-1)})^3 \right\}$$
$$\times \left\{ \sum_{\boldsymbol{\omega} : N_{\boldsymbol{\omega}} > 0} N_{\boldsymbol{\omega}} (\hat{\theta}_{(\cdot)} - \hat{\theta}_{\boldsymbol{\omega}}^{(-1)})^2 \right\}^{-3/2}.$$

These values of the parameters $\hat{z}_0$ and $\hat{a}$ are then used to choose the appropriate percentiles of the bootstrap distribution, using the standard BC_a formulation.

3.4. Some Simulation Results

To compare our method with the standard BIC approach as implemented within *Rcapture* (Rivest and Baillargeon 2019), a simulation study was carried out. The model fitted to the five-list UK data by the stepwise approach was used as a starting point. For this model, the predicted probabilities of each of the 32 possible capture histories (including the empty capture history) were calculated. The overall population size was that estimated by the model fit. The reason for using this model as a basis for a simulation is that it is reasonable to suppose that it will display features likely to be seen when using the methods in the human trafficking context. An example with five lists was used so that the repeated use of the BIC method does not become computationally burdensome.

The population size and the capture history probabilities were regarded as fixed, and were used as parameters for multinomial sampling to create 500 simulated datasets. For each simulation, population estimates and confidence intervals were constructed both using the BIC approach and using the stepwise method we have set out. For the BIC method, multinomial confidence intervals using the routine `closedpCI.t` within *Rcapture* were found; the confidence intervals for the stepwise method were constructed using the BC_a approach. Because the simulations are constructed from a model with known population size, it was possible to assess the accuracy of the estimation. The root mean square error of the estimation was 3057 for the stepwise approach and 5834 for the BIC method. The root mean square errors of the estimate of the log of the population size

were 0.19 and 0.34, respectively, so again, for this example, the stepwise approach has much better performance.

The coverage rate of the estimated confidence intervals was also determined. For the stepwise method using the BCa approach, the nominal 95% confidence interval contained the true value for 90% of the simulations, while the nominal 80% intervals had an actual coverage rate of about 70% (346 out of the 500 replications). While these rates are not perfect, the corresponding observed coverage rates for the methods using routines in *Rcapture* were considerably lower, 61.4% and 42.8%, respectively.

4. Empirical Applications

In this section, our methods are applied to two datasets relating to victims of modern slavery and human trafficking in the USA. Both datasets display the sparseness of overlapping entries typical of data collected in this field. In addition they are also typical of data collected in local regions (rather than entire large countries) in having relatively small counts, with the total number of observed cases in the hundreds and not the thousands. The two datasets, together with those discussed in Section 2.5 are then used to construct a simulation study investigating the appropriate choice of threshold parameter.

4.1. The New Orleans Data

Bales, Murphy, and Silverman (2019) discuss a dataset collated from a number of sources in New Orleans, given in Table 4.

Altogether there are eight lists, and so the full incidence table of observable capture histories, including those combinations for which the actual observed number is zero, has 255 rows. The null capture history, corresponding to the dark figure, of course cannot be observed, and estimating it is the task of the analysis. There are 28 possible pairs of lists, and of these there are 18 nonoverlapping pairs. Using the threshold $p = 0.02$ fits a model including one two-list parameter, indexed by the pair DE. The point estimate of the total population size is 1184. The BC$_a$ bootstrap confidence interval, based on 1000 bootstrap replications, is (717, 1657). If main effects only are chosen (which will be the case for the threshold $p = 0.01$ or smaller), then the resulting model yields a 95% confidence interval of

(644, 1618) with a point estimate of 997. Arguably, with as many as 28 possible two-list parameters, there is some merit in using a smaller threshold.

Because some of the list counts are so small, the effect of combining the four smallest lists into one, to give a five-list version of the data, was also investigated. If this is done, none of the two-list parameters is significant even at the 5% level, and the BC$_a$ confidence interval is (589, 1703) with a point estimate of 1034, a result very close to that yielded by the eight-list data with the smaller threshold. As a further illustration of the issues discussed earlier in the article, and the need to handle nonoverlapping lists in the way we have developed, the *Rcapture* routine `closedpMS.t` was used to fit every possible choice of model with two-list effects. There are 1024 such models, and in only 124 of these was the fit successful without generating a warning. In the majority of cases there was a warning that the asymptotic bias is large.

Return to the full data as an example for the methodology set out in Section 2.7. There are 2^{28} possible models, and 18 nonoverlapping pairs. To check every possible model for existence of the maximum likelihood estimate, there are 2^{18} linear programming problems to solve. This check, which would have been impossible if all 2^{28} models had to be considered explicitly, only takes a few minutes on a standard PC. Neither of the problems identified in Section 2.7 arises for any model for these data.

4.2. The Western Site Data

One of two datasets considered by Farrell et al. (2019) is collated from a number of sources in the Western site of a research study in the USA. The data are given in Table 5.

Altogether there are 5 lists, and so the full incidence table including those combinations for which the observed number is zero has 31 rows. There are 10 possible pairs of lists, and of these there are 2 nonoverlapping pairs. It is very quick to check that all possible models lead to estimates that exist and are identifiable.

The threshold of $p = 0.02$ yields a model including the two-list effect AE, with a point estimate of 2483. The BC$_a$ confidence interval is (1293, 3670).

It should also be noted that the application of `closedpMS.t` to this dataset again generated warnings in more than half of the 1024 possible models. In both this dataset and the New Orleans five-list dataset, warnings were generated among the top 10

Table 4. Victims related to modern slavery and trafficking in New Orleans.

Cases observed only on one list		Cases observed on exactly 2 lists		Cases observed on exactly 3 lists	
List	Number	Lists	Number	Lists	Number
A	25	A&C	1	A&C&G	1
B	5	A&D	2	A&D&E	1
C	70	A&E	1		
D	33	B&F	1		
E	6	C&D	1		
F	6	C&E	1		
G	6	C&G	1		
H	21	D&E	2		
		E&H	1		

NOTE: Numbers of cases on each possible combination of lists, leaving out combinations for which no cases were observed. For reasons of confidentiality the lists are labelled uninformatively.

Table 5. Victims related to human trafficking in the Western site of a research study in the USA.

Cases observed only on one list		Cases observed on exactly 2 lists		Cases observed on exactly 3 lists	
List	Number	Lists	Number	Lists	Number
A	52	A&C	4	A&C&E	1
B	90	A&D	2	B&C&D	1
C	114	A&E	5		
D	45	B&C	6		
E	21	B&D	1		
		D&E	3		

NOTE: Numbers of cases on each possible combination of lists, leaving out combinations for which no cases were observed. For reasons of confidentiality the lists are labeled uninformatively.

models according to the BIC that `closedpMS.t` displays by default, but not, as it happens, by the very top model. For the Netherlands data considered earlier, 6 out of the 10 top models generates a warning.

4.3. Choosing the Threshold: A Simulation Study

To gain insight into the appropriate choice of threshold, a simulation study was carried out. To make this relevant to the context of human trafficking, the models considered are all based on the datasets referenced in this article, in an attempt to ensure that the simulation study is based on datasets that have the kinds of characteristics likely to be encountered. The datasets considered were the UK, Netherlands, New Orleans and Western site data; in the case of the UK, Netherlands and New Orleans data, both the full and the five-list versions were included, giving seven datasets in all. For each of these, four different models were fitted; the "full" model with all two-list effects included, the model based on main effects only, and the models chosen by the method we set out, using thresholds 0.001 and 0.05, to give a more parsimonious and a less restrictive fit. In every case, the model fit gives an estimate of the total population and of the probabilities of all possible capture histories.

For each of these 28 test cases, 1000 realizations of the capture history totals were simulated, by drawing from a multinomial distribution with the given population size and capture history probabilities. Each realization can be conceptualized as an example of a multiple systems sample from a population of known size, with characteristics similar to those likely to be observed in the human trafficking context. For each realization, estimates of the total population were obtained using a range of thresholds (0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1), as well as for the main effects model and the model with all two-list effects included, corresponding to thresholds 0 and 1, respectively. The estimates of the total population size were, as one would expect given the log-linear nature of the modeling, asymmetrically distributed around the true value for each simulation scenario, and a log transformation is appropriate. With this in mind the measure of accuracy used for each of the 28 sets of simulations, for each threshold parameter, was the mean square error of the logarithm of the estimate of the dark figure.

The general level of mean square error varied quite substantially across the 28 models considered. To take account of this variation, for each threshold, the mean, over the 28 models, of the logarithm of the mean square error was calculated to give an overall score for that threshold. The threshold with the minimum score is $p = 0.02$. Further details of the simulation study are given in Section 5 of the supplementary materials to this article.

5. Concluding Remarks

The R software package *SparseMSE* (Chan, Silverman, and Vincent 2019) includes implementations of all the methodology described in this article. In particular, it contains programs to check whether a particular model leads to either of the estimability issues set out in Section 2.6, and it incorporates these checks within a routine to fit any particular model, or to make the model choice using the stepwise procedure described in Section 3.2. It also allows for the possibility of checking all possible models using the approaches discussed in Section 2.7. Full details are given in the package documentation.

To conclude, in this article we have investigated inference for multiple systems estimation using Poisson log-linear models, taking proper account of the possibility that the underlying data tables contain nonoverlapping lists, as commonly arises when the data are collected in the context of studies on modern slavery and human trafficking. We have also set out an approach to model choice and demonstrated the utility and practicality of our approach on real datasets. This area is especially challenging for methodological development because there is no "ground truth" against which methods can be assessed, and frequently there are no details of the data available beyond anonymized list data of the form presented in the tables above. Nevertheless, reliable and stable methods are important for applications in public policy, even if they are conditional on assumptions that it may not be possible to verify.

For simplicity and clarity, the procedure has been discussed and detailed in full for models that only include terms indexed by individual lists and pairs of lists. In principle, the model fitting and inference aspects can easily be extended to consider models including higher order terms, though it seems unlikely that any datasets collected in the contexts of human trafficking would merit this. For example, if a three-list parameter α_{123} were a candidate for inclusion within the model, then the estimate of α_{123} would be $-\infty$ if the three-list overlap N^*_{123} were empty, and to fit the other parameters one would then remove all capture histories including all three lists 1, 2, and 3 from the `glm` stage.

Similarly, another possible extension is to the case where there is covariate information rather than just presence/absence on various lists. As in our main discussion suppose there is a pair (or larger set) of lists whose interaction parameter is in the model but for which no overlapping cases are observed for any value of a covariate. Then the right approach (depending on the exact details of the modeling) would be to set the corresponding interaction parameter to $-\infty$ and then remove various zero cells containing the nonoverlapping set of lists from the fitting procedure for the other parameters including those relevant to covariates.

One possible topic for future research is the combination of our insights with those of International Working Group for Disease Monitoring and Forecasting (1995), which explores the effect of heterogeneity. Some of the approaches suggested in that article may not be available. For example in the human trafficking context we may not be able to stratify the population, nor may the statisticians analyzing the data have any information about the lists themselves to evaluate the possibility of heterogeneity. On the other hand, if one is in a position to implement the proposals, then the possibility of effects of the kind we have explored has to be taken into account.

Supplementary Materials

The supplementary materials for this article contain additional information and details for the simulation carried out to investigate aspects of the asymptotic likelihood theory discussed in Section 2.5; the UK and Netherlands datasets; the statement and proof of a proposition that justifies

the conditions stated in Section 2.6 under which a model is nonidentifiable; and further details of the simulation study, described in Section 4.3, carried out to inform an appropriate choice of p-value threshold for the stepwise algorithm.

Acknowledgments

Funding

References

Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723. [4]

Baillargeon, S., and Rivest, L.-P. (2007), "Rcapture: Loglinear Models for Capture-Recapture in R," *Journal of Statistical Software*, 19, 1–31. [2]

Bales, K., Hesketh, O., and Silverman, B. W. (2015), "Modern Slavery in the UK: How Many Victims?," *Significance*, 12, 16–21. [1]

Bales, K., Murphy, L., and Silverman, B. W. (2019), "How Many Trafficked People Are There in Greater New Orleans? Lessons in Measurement," *Journal of Human Trafficking*, DOI: 10.1080/23322705.2019.1634936. [1,2,8]

Bird, S. M., and King, R. (2018), "Multiple Systems Estimation (or Capture-Recapture Estimation) to Inform Public Policy," *Annual Review of Statistics and Its Application*, 5, 95–118. [1,2]

Chan, L., Silverman, B. W., and Vincent, K. (2019), "SparseMSE: Multiple Systems Estimation for Sparse Capture Data," R Package Version 2.0.1. [2,9]

Cormack, R. M. (1989), "Log-Linear Models for Capture–Recapture," *Biometrics*, 45, 395–413. [2,3]

——— (1992), "Interval Estimation for Mark-Recapture Studies of Closed Populations," *Biometrics*, 48, 567–576. [2]

Cruyff, M., van Dijk, J., and van der Heijden, P. G. M. (2017), "The Challenge of Counting Victims of Human Trafficking: Not on the Record: A Multiple Systems Estimation of the Numbers of Human Trafficking Victims in the Netherlands in 2010–2015 by Year, Age, Gender, and Type of Exploitation," *CHANCE*, 30, 41–49. [1]

DiCiccio, T. J., and Efron, B. (1996), "Bootstrap Confidence Intervals," *Statistical Science*, 11, 189–228. [2,7]

Far, S. S., Papathomas, M., and King, R. (2019), "Parameter Redundancy and the Existence of Maximum Likelihood Estimates in Log-Linear Models," *Statistica Sinica*, DOI: 10.5705/ss.202018.0100. [5]

Farrell, A., Dank, M., Kfafian, M., Lockwood, S., Pfeffer, R., Hughes, A., and Vincent, K. (2019), "Capturing Human Trafficking Victimization Through Crime Reporting," Technical Report 2015-VF-GX-0105, National Institute of Justice. Final Summary Report, available at *https://www.ncjrs.gov/pdffiles1/nij/grants/252520.pdf*. [2,8]

Fienberg, S. E., and Rinaldo, A. (2012a), "Maximum Likelihood Estimation in Log-Linear Models," *The Annals of Statistics*, 40, 996–1023. [2,3,5]

——— (2012b), "Maximum Likelihood Estimation in Log-Linear Models: Supplementary Material," Technical Report, Carnegie Mellon University. [5]

Home Office (2014), *Modern Slavery Strategy*, London, UK: HM Government, available at *https://www.gov.uk/government/publications/modern-slavery-strategy*. [1,4]

International Working Group for Disease Monitoring and Forecasting (1995), "Capture-Recapture and Multiple-Record Systems Estimation II: History and Theoretical Development," *American Journal of Epidemiology*, 142, 1047–1058. [9]

King, R., Bird, S. M., Overstall, A. M., Hay, G., and Hutchinson, S. J. (2013), "Injecting Drug Users in Scotland, 2006: Number, Demography, and Opiate-Related Death-Rates," *Addiction Research and Theory*, 21, 235–246. [1]

Petersen, C. (1896), "The Yearly Immigration of Young Plaice Into the Limfjord From the German Sea," *Report of the Danish Biological Station*, 6, 5–84. [1]

R Core Team (2016), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [2]

Rivest, L.-P., and Baillargeon, S. (2019), "Rcapture: Loglinear Models for Capture-Recapture Experiments," R Package Version 1.4-3. [2,4,7]

Rivest, L.-P., and Daigle, G. (2004), "Loglinear Models for the Robust Design in Mark-Recapture Experiments," *Biometrics*, 60, 100–107. [2]

Sandland, R. L., and Cormack, R. M. (1984), "Statistical Inference for Poisson and Multinomial Models for Capture–Recapture Experiments," *Biometrika*, 71, 27–33. [2,7]

Schwarz, C. J., and Seber, G. A. F. (1999), "Estimating Animal Abundance: Review III," *Statistical Science*, 14, 427–456. [1]

Silverman, B. W. (2020), "Multiple-Systems Analysis for the Quantification of Modern Slavery: Classical and Bayesian Approaches" (with discussion), *Journal of the Royal Statistical Society*, Series A, 183. [4]

van Dijk, J. J., Cruyff, M., van der Heijden, P. G. M., and Kragten-Heerdink, S. L. J. (2017), *Monitoring Target 16.2 of the United Nations' Sustainable Development Goals; A Multiple Systems Estimation of the Numbers of Presumed Human Trafficking Victims in the Netherlands in 2010–2015 by Year, Age, Gender, Form of Exploitation and Nationality*, Vienna, Austria: United Nations Office on Drugs and Crime. [4]

Wilks, S. S. (1938), "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses," *The Annals of Mathematical Statistics*, 9, 60–62. [4]

Williams, B. K., Nichols, J. D., and Conroy, M. (2002), *The Analysis and Management of Animal Populations*, San Diego, CA: Academic Press. [1]