Multiple Systems Estimation for Sparse Capture Data: Inferential Challenges when there are Non-Overlapping Lists

Lax Chan, Bernard W. Silverman and Kyle Vincent
Rights Lab, University of Nottingham, U.K.

April 26, 2019

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 and maximum likelihood inference of the parameters. 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. We apply the strategy to two empirical data sets 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.

Keywords: Human trafficking; Log-linear models; Mark-recapture; Model identifiability; Model selection; Modern slavery; Poisson regression modeling.

1 Introduction

Multiple systems estimation, a generalization of the classic mark-recapture approach (Petersen, 1896; Schwarz and Seber, 1999), is a class of methods that are used to estimate the size of hard-to-reach populations in many different contexts, including, in recent years, those comprised of human trafficking or slavery victims. The method is typically applied to wildlife populations (Williams et al., 2002) and more recently to hidden populations like those comprised of injection drug users (King et al., 2013). In the administrative or law enforcement context, multiple systems estimation is an approach to read across from lists of observed or identified individuals from the study population to gain an estimate of the total population of interest; original contributions in the subject have been made by Bales et al. (2015) and Cruyff et al. (2017). Typically, a mathematical model is posited for the pattern of incidences across the lists, and the "dark figure" of unobserved cases is estimated. A discussion of the history of the methods and a survey of a range of applications is provided, for example, by Bird and King (2018).

Multiple systems estimation provides a method to quantify the number of victims in a study population, including those that are not directly observed or detected. The method therefore plays an especially important role as it assists with policy making decisions to help combat human trafficking and slavery. For example, as set out in Bales et al. (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 specific challenge posed by data on human trafficking is that it is not uncommon for there to be sparse overlap between the observed administrative lists. This is in contrast with the applications such as wildlife populations, where the researcher may have a degree of control over sample sizes and can continue capturing from the study population until sufficient overlap is observed between the capture occasions. Of course, in the original mark-recapture setup where there are only two capture occasions, it is the overlap between the lists which allows the inference to be conducted at all.

However, in the human trafficking context, 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 may be for several reasons: there may be a genuine structural reason why two particular lists cannot overlap; there may be negative correlation between lists; or it may simply be that for a relatively small sample size and two lists with small capture probabilities there do not happen to be any cases which are on both lists. In this area, there is as yet limited understanding of data and of mechanisms, and furthermore data are often highly anonymised for reasons of confidentiality and security, to the point that we may not know anything about a list other than an anonymised label. Hence, there may be no further information available as to why no cases are observed in common between two lists.

We approach inference via Poisson regression modeling applied to counts of individuals that are observed on each possible combination of the lists, with model parameters that correspond to each combination of lists. We use this approach since it is a well-known technique that allows one to model list interactions, and since within the multiple systems estimation context inference is naturally based on categorical data for which a log-linear approach is ideal. The standard log-linear approach is set out by Cormack (1989), Rivest and Daigle (2004) and Bird and King (2018), among others, and implemented in Baillargeon and Rivest (2007) and Baillargeon and Rivest (2012). However, as Fienberg and Rinaldo (2012a) discuss in a much more general context, contingency tables with zero entries may lead to cases where the maximum likelihood estimate of the model parameters does not exist or is not identifiable. In the context of multiple systems estimation, empty overlaps between lists require careful treatment for these reasons. The primary objective of this paper is to introduce inferential procedures and computational implementations that explicitly handle this case.

We first of all develop a method which fits a model stably to data of this kind, 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. The assumptions that underlie information-theoretic approaches do not hold, and so a stepwise approach is used, motivated and illustrated by data sets based on human trafficking victims in the New Orleans area (Bales et al., 2018) and the Western site of a research study in the USA (Farrell et al., 2019). We conduct our analyses in the R programming language (R Core Team, 2016), and have developed an accompanying R software package `SparseMSE` (Chan et al., 2019). The package allows readers to implement the methodology on their own data as well as to reproduce the results presented in the paper.

The paper is organized as follows. Section 2 outlines the model and gives the notation and likelihood setup, and also details specific issues concerning the existence and identifiability of maximum likelihood estimates. Section 3 presents the model-selection routine and corresponding inference procedure. Section 4 presents the results from the two empirical applications. A comment on the R package and some concluding remarks are made in Section 5.

2 The Model

In this section, we define notation and then define and explore the model. The discussion leads to an algorithmic approach which allows the correct and stable calculations to be carried out.

2.1 Notation and Definitions

2.1.1 Capture Histories

Suppose we have t capture occasions, or lists, on which members of the population can occur. An individual's *capture history* is the set $\boldsymbol{\omega}$ of lists on which the individual is actually observed, or captured. The capture history is a subset of $\{1, 2, \ldots, t\}$. The *order* of a capture history is defined to be the number of captures in the set. For each capture history $\boldsymbol{\omega}$ define $N_{\boldsymbol{\omega}}$ to be the number of individuals in our population with exactly that capture history.

A particular capture history, with order 0, is the *null capture history* $\emptyset$. The quantity $N_\emptyset$ is the so-called *dark figure* of individuals which are not captured on any list, and therefore cannot be observed. Including the null capture history there are 2^t possible capture histories; the data we will have are the $2^t - 1$ observed values $\{N_{\boldsymbol{\omega}} : \boldsymbol{\omega} \neq \emptyset\}$, which we will also write as $\mathbf{N}$.

When $\boldsymbol{\omega}$ is a suffix the braces are usually omitted and the members of the history simply 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} if $t < 10$, is the number of individuals which are observed on both lists 1 and 3 but not on lists 2 or 4. While we will collapse $N_{\{i\}}$ to N_i and $N_{\{i,j\}}$ to N_{ij}, the dark figure will always be written $N_\emptyset$, because the letter N is reserved for the total population size including both the observed cases and the dark figure.

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.

2.1.2 Further Notation

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 for which $\omega_i = 1$, regardless of whether they do or do not appear on other lists. For example, N^*_1 is the total number of individuals on list 1, while N_1 is the number of individuals that are on list 1 but none of the other lists. To move to pairs of lists, N^*_{12} is the total overlap between 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, … . Finally note that 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 number of individuals actually captured at some point in the study.

We will call $\{i, j\}$ a *non-overlapping pair* of lists if $N^*_{ij} = 0$, so that no individual appears on both lists. The main objective of this paper is to deal properly with data that contains such non-overlapping pairs.

2.2 The Poisson Loglinear Model

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

$$N_{\boldsymbol{\omega}} \sim \text{Poisson}(\mu_{\boldsymbol{\omega}})$$

where

$$\log \mu_{\boldsymbol{\omega}} = \sum_{\boldsymbol{\theta} \subseteq \boldsymbol{\omega}} \alpha_{\boldsymbol{\theta}}$$

for certain parameters $\alpha_{\boldsymbol{\theta}}$ indexed by the possible capture histories. It should be noted that capture histories are used in two different ways, firstly to index the observed data, and secondly 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.

The parameter $\alpha_\emptyset$ indexed by the null capture history will be written more simply as α. Thus, for example, the dark figure has expected value $\exp\alpha = \exp\alpha_\emptyset$, while the expected value of N_{13} is $\exp(\alpha + \alpha_1 + \alpha_3 + \alpha_{13})$.

Altogether, there are 2^t parameters $\alpha_{\boldsymbol{\theta}}$, corresponding to the 2^t capture histories including the null capture history. But the number of cases $N_\emptyset$ for the null capture history is not observed, and so there are only $2^t - 1$ data points from which to estimate the parameters; without placing constraints on the $\alpha_{\boldsymbol{\theta}}$ parameters, the model is not identifiable.

As Cormack (1989) sets 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 interaction coefficients indexed by third- or higher-order histories to zero, and we will do this throughout. Even if all the second-order coefficients are included, the number of parameters to be estimated is $1 + t + \frac{1}{2}t(t-1) \leq 2^t - 1$ provided $t \geq 3$. The question of model choice then reduces to deciding which of the second-order interactions to include, and will be discussed further in Section 3 below. For any particular set of second-order interactions, the parameter estimation can be put into a standard generalized linear model formulation.

A consequence of the definition is that it is also the case that, for each $\boldsymbol{\omega}$,

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

Unlike the $N_{\boldsymbol{\omega}}$, the $N^*_{\boldsymbol{\omega}}$ are not independent of one another. For example, if capture histories $\boldsymbol{\omega}$ and $\boldsymbol{\psi}$ share one or more 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 non-overlapping pairs of lists, it will be helpful to derive some properties of the log 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 set of parameters to be estimated. Note that Θ will always contain $\emptyset$.

Up to an additive constant depending only on the data, the log likelihood is given by

$$\ell(\boldsymbol{\alpha}|\mathbf{N}) = \sum_{\boldsymbol{\omega} \neq \emptyset} \{N_{\boldsymbol{\omega}} \log(\mu_{\boldsymbol{\omega}}) - \mu_{\boldsymbol{\omega}}\}. \tag{2}$$

Consider the first term, substituting the definition of the model, reversing the order of summation, and then substituting the definition (1), to obtain

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

Turning to the other term in the log likelihood,

$$-\sum_{\boldsymbol{\omega}\neq\emptyset} \mu_{\boldsymbol{\omega}} = \sum_{\boldsymbol{\omega}\neq\emptyset}\{-\exp\left[\sum_{\boldsymbol{\theta}\subseteq\boldsymbol{\omega},\boldsymbol{\theta}\in\Theta} \alpha_{\boldsymbol{\theta}}\right]\} = C(\boldsymbol{\alpha}), \qquad (4)$$

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 concave function because it 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) point out and explore in a much more general and abstract context, and as we shall see below, it is not necessarily the case that there is a unique maximum likelihood estimate of $\boldsymbol{\alpha}$ nor that such an estimate exists at all.

These expressions for the components of the log likelihood function demonstrate the following, which will be useful when we come to discuss model choice below:

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 Non-Overlapping Pairs

Suppose that $\{i, j\}$ is a non-overlapping 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; this section gives an elementary recapitulation of some of their results cast into our specific framework. It follows from (3) that the first term of the log likelihood does not depend on α_{ij}, because the coefficient of α_{ij} is zero. So the maximum likelihood estimate of α_{ij} will be obtained by maximizing the function $C(\boldsymbol{\alpha})$. Because, as already noted, $C(\boldsymbol{\alpha})$ 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 the existing software packages yield errors or warnings if there are non-overlapping 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} yields legitimate values for the actual Poisson parameters, giving the value zero for $\mu_{\boldsymbol{\omega}}$ for all $\boldsymbol{\omega} \supseteq \{i, j\}$, provided we regard 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 $\boldsymbol{\alpha}^{\dagger}_{ij}$ for the vector of parameters with α_{ij} excluded,

$$\ell(\boldsymbol{\alpha}^{\dagger}_{ij}|\mathbf{N}, \alpha_{ij} = -\infty) = \sum_{\boldsymbol{\omega}\neq\emptyset, \boldsymbol{\omega}\not\supseteq\{i,j\}} \{N_{\boldsymbol{\omega}} \log(\mu_{\boldsymbol{\omega}}) - \mu_{\boldsymbol{\omega}}\}. \tag{5}$$

This is exactly the Poisson log likelihood based on all the observations except those for the 2^{t-2} capture histories which include both i and j. If there is more than one non-overlapping pair in Θ, the same calculations can be carried out for each pair, leading to the following algorithm:

1. 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 list of parameters to be estimated. Let $\Theta^\dagger$ be the set of indices of parameters in Θ that have not been removed.

2. For each such $\{i, j\}$ remove from consideration all $N_{\boldsymbol{\omega}}$ for which $\boldsymbol{\omega} \supseteq \{i, j\}$, regarding them as structural zeroes. Let $\Omega^\dagger$ be the set of capture histories whose counts $N_{\boldsymbol{\omega}}$ remain (not including the null capture history).

3. Use the standard generalized model approach to estimate the parameters with indices in $\Theta^\dagger$ from the observed counts of the capture histories in $\Omega^\dagger$.

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 Estimability of the Population Size

There are two separate estimability issues that 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 the maximum likelihood estimate does not exist, even if we take account of non-overlapping pairs in the way set out above. Fienberg and Rinaldo (2012b) show that existence of the estimate can be checked by solving a linear programming problem. 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 $\Omega^{\dagger}$, so that $\mathbf{A}_{\boldsymbol{\omega}\boldsymbol{\theta}} = 1$ 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 maximizing s over all scalars s and all real vectors $\mathbf{x} = (x_{\boldsymbol{\omega}}, \boldsymbol{\omega} \in \Omega^{\dagger})$ satisfying the constraints

$$\begin{aligned} \mathbf{A}^T\mathbf{x} &= \mathbf{t} \text{ and} \\ x_{\boldsymbol{\omega}} - s &\geq 0 \text{ for all } \boldsymbol{\omega} \in \Omega^{\dagger}. \end{aligned} \tag{6}$$

A necessary and sufficient condition for a maximum likelihood estimate to exist is that the maximizing value 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 to the constraints (6). Hence the maximizing value of s will be at least the minimum of the observed $N_{\boldsymbol{\omega}}$ over $\Omega^{\dagger}$. In the non-sparse 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 to consider is that although the likelihood can be maximized, the parameters are unidentifiable, in the sense that any parameter set in a particular non-trivial affine subspace of the parameter space $\Theta^{\dagger}$ will maximize the likelihood. In particular the parameter which estimates the dark figure will not be estimable, and hence nor will the population size. The model will be identifiable if and only if $\mathbf{A}$ is of full column rank. We show in Section 2.6.2 that non-identifiability 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 non-overlapping 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 unidenfiability directly, more often arbitrarily removing one or more of the parameters. Unless every possible capture

Table 1: An artificial data set with three lists.

A	B	C	count
×			40
	×		30
		×	20
×	×		6

history is actually represented in the observed data, therefore, it is important to check that a potential model is identifiable, and gives a strictly positive value for the linear programming problem. If it fails on either count it should be ruled out. These checks incur only a small computational overhead.

The simple example given in Table 1 shows that there is not necessarily any clear hierarchical relationship between models that fail on the existence criteria. Within our framework there are eight possible choices of the pairwise effects to include in the model. The results obtained from the checks are as follows. If all three interactions are included, then the model is not identifiable but nevertheless the linear program yields a strictly positive value. However, if the model contains AB either alone or in conjunction with one of AC and BC, then the linear program result is zero, so the MLE does not exist. If only main effects are considered, or if either or both of AC and BC, but not AB are included, then the model passes both tests. There is further discussion of the possible effect of adding and removing parameters in Section 2.6.1 below.

2.6 Checking all models

Given a particular data set, it is useful in certain contexts to check that the estimates satisfy the existence and identifiability criteria, no matter which pairwise terms are included in the model. In this section, we set out the justification for algorithmic approaches which allow this to be done 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 parameters $\alpha_\emptyset$ and 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.

2.6.1 The Linear Program problem

First consider the Fienberg-Rinaldo conditions for the extended maximum likelihood estimates to exist. 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 model. Now consider the effect of removing $\{i, j\}$ from the model. Because $\{i, j\}$ is an overlapping pair, this will not change the set $\Omega^\dagger$, but it will remove $\{i, j\}$ from $\Theta^\dagger$. In the linear programming problem (6), 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 can only increase. This means that 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 check whether all models satisfy the conditions for estimates to exist, it is only necessary to test parameter sets Θ that include all overlapping pairs and a subset (possibly empty) of the non-overlapping pairs. If there are M non-overlapping 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 data sets than if all models have to be considered.

Furthermore, in the event that there are models for which the estimate does not exist, this approach can be extended to find a list of all such models efficiently. Let Θ_2^{non} be the set of all non-overlapping pairs and Θ_2^{over} the set of overlapping pairs. Suppose that the search over subsets of Θ_2^{non} yields a model Θ, containing the subset $\tilde{\Theta}_2^{\text{non}}$ of Θ_2^{non} but all of Θ_2^{over}, for which the maximum likelihood estimate does not exist. We then perform a hierarchical search 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 one of the resulting models still 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.

2.6.2 Identifiability

Now consider the question of identifiability. We show that the only model that can fail to be identifiable is the case where *all* the two-list terms are included in the parameter set Θ. Furthermore there is a simple algorithm to determine whether or not that model is also identifiable. All models which include only a proper subset of all the two-list terms necessarily yield a matrix A of full column rank.

Recall that $\mathbf{A}_{\boldsymbol{\omega}\boldsymbol{\theta}} = 1$ if $\boldsymbol{\theta} \subseteq \boldsymbol{\omega}$ and 0 otherwise. Suppose that $\boldsymbol{\lambda} = (\lambda_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta^\dagger)$ is a vector such that $\mathbf{A}\boldsymbol{\lambda} = 0$. Then

$$\sum_{\boldsymbol{\theta}\in\Theta^\dagger} \mathbf{A}_{\boldsymbol{\omega}\boldsymbol{\theta}}\lambda_{\boldsymbol{\theta}} = 0. \tag{7}$$

By its definition, the set $\Omega^\dagger$ includes all capture histories $\{i\}$ of order 1. From (7) and the definition of $\mathbf{A}$, we have, for each i,

$$\sum_{\boldsymbol{\theta}\in\Theta^\dagger} \mathbf{A}_{\{i\}\boldsymbol{\theta}}\lambda_{\boldsymbol{\theta}} = \lambda_\emptyset + \lambda_i = 0, \tag{8}$$

so $\lambda_i = -\lambda_\emptyset$ for all i. Furthermore, for any pair of lists $\{i,j\} \in \Theta^\dagger$, $\{i,j\}$ must be in $\Omega^\dagger$, and

$$\sum_{\boldsymbol{\theta}\in\Theta^\dagger} \mathbf{A}_{\{i,j\}\boldsymbol{\theta}}\lambda_{\boldsymbol{\theta}} = \lambda_\emptyset + \lambda_i + \lambda_j + \lambda_{ij} = 0. \tag{9}$$

so that $\lambda_{ij} = \lambda_\emptyset$ for all $\{i,j\} \in \Theta^\dagger$.

Now suppose there is a pair of lists, without loss of generality $\{1,2\}$, that is not in the parameter set Θ. The corresponding capture history $\{1,2\}$ will therefore not have been removed from $\Omega^\dagger$. More concisely, we have $\{1,2\} \in \Omega^\dagger$, but $\{1,2\} \notin \Theta^\dagger$, since $\Theta^\dagger \subseteq \Theta$. Hence we have

$$\sum_{\boldsymbol{\theta}\in\Theta^\dagger} \mathbf{A}_{\{1,2\}\boldsymbol{\theta}}\lambda_{\boldsymbol{\theta}} = \lambda_\emptyset + \lambda_1 + \lambda_2 = 0.$$

Substituting equation (8) for $i = 1$ and $i = 2$ shows that $\lambda_\emptyset = 0$ and hence $\lambda_i = \lambda_{ij} = 0$ for all i and for all $\{i,j\} \in \Theta^\dagger$. Thus $\lambda_{\boldsymbol{\theta}} = 0$ for all $\boldsymbol{\theta} \in \Theta^\dagger$, and so $\boldsymbol{\lambda} = 0$. Hence the columns of $\mathbf{A}$ are linearly independent and $\mathbf{A}$ is of full column rank.

Thus the only possible nonidentifiable model is the one which includes parameters for all the capture histories of order 2. Consider that model, and suppose as above, that $\mathbf{A}\boldsymbol{\lambda} = 0$.

Suppose, now, that there are three lists i, j and k such that none of the pairs $\{i,j\}$, $\{j,k\}$ and $\{i,k\}$ are non-overlapping. Hence $\{i,j\}, \{i,k\}$ and $\{j,k\}$ are all in $\Theta^\dagger$. The capture history $\{i,j,k\}$ will be in $\Omega^\dagger$, and so

$$0 = \sum_{\boldsymbol{\theta}\in\Theta^\dagger} \mathbf{A}_{\{i,j,k\}\boldsymbol{\theta}}\lambda_{\boldsymbol{\theta}} = \lambda_\emptyset + \lambda_i + \lambda_j + \lambda_k + \lambda_{ij} + \lambda_{ik} + \lambda_{jk} = \lambda_\emptyset,$$

making use of equations (8) and (9). As before this will imply that $\lambda_{\boldsymbol{\theta}} = 0$ for all $\boldsymbol{\theta} \in \Theta^\dagger$, so that $\boldsymbol{\lambda} = 0$, and hence the model will be identifiable.

Now suppose, conversely, that $\Theta^\dagger$ contains no such triple pairs, so that every set of three lists contain at least one non-overlapping pair. The set $\Omega^\dagger$ will contain no capture histories of order 3 or above. Now, define $\lambda_\emptyset = 1, \lambda_i = -1$ for all i and $\lambda_{ij} = 1$ for all $\{i, j\} \in \Theta^\dagger$. Then every element of $\mathbf{A}\boldsymbol{\lambda}$ will be calculated as in one of the two equations (8) or (9). Hence $\mathbf{A}\boldsymbol{\lambda} = 0$ for a non-zero vector $\boldsymbol{\lambda}$ and so the model is, in this case, not identifiable.

To sum up, for the model containing all pairs $\{i, j\}$, $\mathbf{A}$ is of full column rank if and only if there is at least one set of three lists $\{i, j, k\}$ that contains no non-overlapping pairs. To check this simply, define the matrix J by $J_{ij} = 1$ if $\{i, j\}$ is an overlapping pair, and zero otherwise, with all $J_{ii} = 0$. The model will be identifiable if and only if $\text{trace}(J^3) > 0$. To see this, note that $\text{trace}(J^3) = \sum_{i,j,k} J_{ij}J_{jk}J_{ki}$. The terms in this sum are all zero or one, and the trace will be strictly positive if and only if there is at least one non-zero term $J_{ij}J_{jk}J_{ki}$, in other words if $\{i, j, k\}$ contains no non-overlapping pairs.

3 Inference and Model Choice

3.1 Calculating Significance

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

$$\hat{\mu}_{\boldsymbol{\omega}}[\Theta] = \exp\left(\sum_{\boldsymbol{\theta} \subseteq \boldsymbol{\omega}, \boldsymbol{\theta} \in \Theta} \hat{\alpha}_{\boldsymbol{\theta}}\right) \tag{10}$$

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

$$\hat{\mu}^*_{\boldsymbol{\omega}}[\Theta] = \sum_{\boldsymbol{\psi} \supseteq \boldsymbol{\omega}, \boldsymbol{\psi} \neq \emptyset} \hat{\mu}_{\boldsymbol{\psi}}[\Theta]. \tag{11}$$

First, consider how to deal with non-overlapping pairs within the data (note that we only consider a model with up to two-factor interactions). Suppose the model being fitted corresponds to a set of parameters indexed by Θ, and that for some $\boldsymbol{\theta} \in \Theta$ that $N^*_{\boldsymbol{\theta}} = 0$. Should we actually include $\boldsymbol{\theta}$ in the model? To answer the question we test the null hypothesis that $\alpha_{\boldsymbol{\theta}} = 0$, which is equivalent to saying that $\boldsymbol{\theta}$ is not in the model. The natural statistic on which to base a test is $N^*_{\boldsymbol{\theta}}$, because of the results on sufficient statistics in Section 2.3.

Now proceed as follows:

1. Fit the model leaving out the parameter $\alpha_{\boldsymbol{\theta}}$, in other words using just the parameter set $\Theta \setminus \boldsymbol{\theta}$.

2. For the resulting fitted model, find the estimate $\hat{\mu}^*_{\boldsymbol{\theta}}[\Theta \setminus \boldsymbol{\theta}]$, as defined in (10) and (11).

3. Because the probability of observing zero in a Poisson distribution with parameter λ is $e^{-\lambda}$, the estimated parameter has p-value $\exp(-\hat{\mu}^*_{\boldsymbol{\theta}}[\Theta \setminus \boldsymbol{\theta}])$. This is the probability of observing $N^*_{\boldsymbol{\theta}} = 0$ in the model defined by $\Theta \setminus \boldsymbol{\theta}$.

In fact, this approach can be generalised to construct a p-value for *any* interaction $\boldsymbol{\theta} \in \Theta$ whether or not $N^*_{\boldsymbol{\theta}} = 0$. Again calculating $\hat{\mu}^*_{\boldsymbol{\theta}}[\Theta \setminus \boldsymbol{\theta}]$ from the maximum likelihood estimate over $\Theta \setminus \boldsymbol{\theta}$, 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 $F(n, \lambda)$ is the lower tail probability that a Poiss(λ) random variable X satisfies $X \leq n$, while $\tilde{F}(n, \lambda)$ is the probability that $X \geq n$.

3.2 Model Fitting

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

- Step 1: Set a threshold value for the p-value.

- Step 2: Fit the model with the main effects parameters only.

- Step 3: Consider in turn each interaction not already added to the model, and check that adding it to the model would not lead to a non-existent or unidentifiable estimate.

- Step 4: Among those interactions that pass the checks, find the one with the smallest p-value.

- Step 5: If that p-value is less than or equal to the given threshold, add the interaction to the model, and go back to Step 3. If the p-value is greater than the threshold, finish.

Notice that the method is akin to forward stepwise regression since parameters are added stepwise to the model. It remains to choose the threshold p-value, and there are two considerations to bear in mind. Firstly, because we are considering multiple candidates to add to the model, it is appropriate to apply a Bonferroni correction to any standard p-value we might use. So, for example, if there are 8 lists and hence 28 possible two-list interactions, even if we started with the conventional $p = 0.05$, the appropriate threshold would be $0.05/28 = 0.0018$. The second consideration is that we might wish to take a relatively conservative approach aiming at parsimonious models unless the data convincingly argue otherwise. For both these reasons, we suggest using a value of $p = 0.001$ but to explore the sensitivity of the result to adjusting the parameter.

One of the standard approaches in the literature is to use approaches based on information criteria for model choice. The justification of information criteria is based on the asymptotic theory of maximum likelihood estimates (Akaike, 1974) which breaks down when dealing with Poisson parameters close to zero.

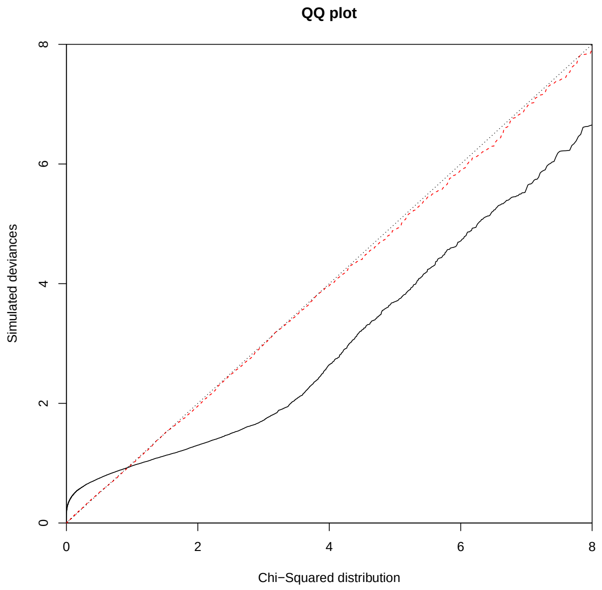
A simulation example which demonstrates this in the multiple systems estimation context is presented in Figure 1. The simulation is carried out using three lists, with specified capture probabilities. Captures on the various lists are assumed to be independent. Two models are used, one with capture probabilities 0.01, 0.04 and 0.2 and the other with capture probabilities 0.3 for each list. The first model is chosen to be somewhat more typical of the sparse capture case, of the kind which often occurs in the human trafficking context, while the second is reminiscent of a more classical mark-recapture study.

The probability of an individual having each possible capture history $\boldsymbol{\omega}$, is first evaluated. Then these probabilities are multiplied by 1000 and, for each simulation replicate, Poisson random values with expectations equal to these values are generated to give a full set of observed capture histories; together with the null capture history the expected number of counts (population size) is equal to 1000. The correct model for these data includes main effects only, so inference was carried out both for that model, and for the model with the addition of an interaction effect between the first two lists. The reduction in deviance between the two models was determined. Ten thousand simulations were carried out.

In line with the standard asymptotic theory, the QQ-plots show that the χ^2_1 distribution fits the observed deviances well for the "classic" model.

However the fit for the sparse model is not good. This demonstrates why information criteria cannot be relied on for fitting models in the sparse context.

Figure 1: QQ plots for sparse model (solid line) and for classic model (dashed line) against quantiles of the χ^2_1 distribution. The dotted line $x = y$ is followed closely by the QQ plot for the classic model.



4 Empirical Applications

In this section, our methods are applied to two data sets relating to victims of modern slavery and human trafficking in the USA. Both data sets display the sparseness of overlapping entries typical of data collected in this field.

4.1 The New Orleans Data

Bales et al. (2018) discuss a data set collated from a number of sources in New Orleans, given in Table 2.

Table 2: Victims related to modern slavery and trafficking in New Orleans. 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 anonymised.

A	B	C	D	E	F	G	H	count
×								25
	×							5
		×						70
			×					33
				×				6
					×			6
						×		6
							×	21
×		×						1
×			×					2
×				×				1
	×				×			1
		×	×					1
		×		×				1
		×				×		1
			×	×				2
				×			×	1
×		×				×		1
×			×	×				1

Altogether there are 8 lists, and so the full incidence table including those combinations for which the observed number is zero has 255 rows. There are 28 possible pairs of lists, and of these there are 18 non-overlapping pairs. The results are relatively insensitive to the choice of threshold p-value; even a threshold as large as $p = 0.01$ does not detect any significant interactions and fits a model based on main effects only, and with 28 possible interactions it is clearly inappropriate to use any larger threshold. The resulting model yields a 95% confidence interval of (644, 1618) with a point estimate of 997. 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, then the confidence interval is (658, 1709) with a point estimate of 1034, and so the results are essentially the same. For the five list data, none of the interactions was significant even at the 5% level.

Even with 18 non-overlapping pairs, it is easily feasible to check every possible model for existence of the maximum likelihood estimate, using the algorithm set out in Section 2.6.1. Even with 2^{18} linear programming problems to solve, the full check only takes a few minutes on a standard PC, and it shows that neither of the problems identified in Section 2.6 arises for any model for these data.

4.2 The Western Site Data

One of two data sets 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 3.

Table 3: Victims related to human trafficking in the Western site of a research study in the USA. 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 anonymised.

A	B	C	D	E	count
×					52
	×				90
		×			114
			×		45
				×	21
×		×			4
×			×		2
×				×	5
	×	×			6
	×		×		1
			×	×	3
×		×		×	1
	×	×	×		1

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 non-overlapping pairs. Applying the Bonferroni correction to the conventional 0.05 would yield a value of $0.05/10 = 0.005$ for the threshold p-value. In fact, using a larger threshold of $p = 0.01$ would give the same result as that based on the Bonferroni correction; only a pairwise interaction of lists A and E is included to fit a model. The resulting model yields a 95% confidence interval of $(1657, 3830)$ with a point estimate of 2484.

5 Concluding Remarks

The R software package `SparseMSE` (Chan et al., 2019) includes implementations of all the methodology described in this paper. In particular, it contains programs to check whether a particular model leads to either of the estimability issues set out in Section 2.5 above, 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 2.6. Full details are given in the package documentation.

To conclude, in this paper 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 non-overlapping 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 data sets. 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 anonymised 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 which consider up to two-way interaction terms. In principle, the model checking and inference aspects can easily be extended to consider models based on higher-order interaction terms, but it seems unlikely that any data sets collected in the contexts of our current interest would merit this.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19,** 716–723.

Baillargeon, S. and Rivest, L.-P. (2007). Rcapture: loglinear models for capture-recapture in R. *Journal of Statistical Software* **19,** 1–31.

Baillargeon, S. and Rivest, L.-P. (2012). *Rcapture: Loglinear Models for Capture-Recapture Experiments.* R package version 1.3-1.

Bales, K., Hesketh, O., and Silverman, B. W. (2015). Modern slavery in the UK: How many victims? *Significance* **12,** 16–21.

Bales, K., Murphy, L., and Silverman, B. W. (2018). How many trafficked people are there in New Orleans? Lessons in measurement. Preprint, available from `https://tinyurl.com/ybfb9tg6`.

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.

Chan, L., Silverman, B. W., and Vincent, K. (2019). *SparseMSE: Multiple systems estimation for sparse capture data.* R package, available from `https://CRAN.R-project.org/package=SparseMSE`.

Cormack, R. M. (1989). Log-linear models for capture-recapture. *Biometrics* **45,** 395–413.

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.

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 from `https://www.ncjrs.gov/pdffiles1/nij/grants/252520.pdf`.

Fienberg, S. E. and Rinaldo, A. (2012a). Maximum likelihood estimation in log-linear models. *Ann. Statist.* **40,** 996–1023.

Fienberg, S. E. and Rinaldo, A. (2012b). Maximum likelihood estimation in log-linear models: supplementary material. Technical report, Carnegie Mellon Univ.

Home Office (2014). Modern Slavery Strategy. `https://www.gov.uk/government/publications/modern-slavery-strategy`.

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.

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.

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

Rivest, L.-P. and Daigle, G. (2004). Loglinear models for the robust design in mark-recapture experiments. *Biometrics* **60,** 100–107.

Schwarz, C. J. and Seber, G. A. F. (1999). Estimating animal abundance: Review III. *Statistical Science* **14,** 427–456.

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