Statistics and Computing (2024) 34:44
https://doi.org/10.1007/s11222-023-10346-9

ORIGINAL PAPER

Bootstrapping multiple systems estimates to account for model selection

Bernard W. Silverman[1] · Lax Chan[2] · Kyle Vincent[3]

Received: 21 June 2023 / Accepted: 17 October 2023
© Crown 2023

Abstract
Multiple systems estimation using a Poisson loglinear model is a standard approach to quantifying hidden populations where data sources are based on lists of known cases. Information criteria are often used for selecting between the large number of possible models. Confidence intervals are often reported conditional on the model selected, providing an over-optimistic impression of estimation accuracy. A bootstrap approach is a natural way to account for the model selection. However, because the model selection step has to be carried out for every bootstrap replication, there may be a high or even prohibitive computational burden. We explore the merit of modifying the model selection procedure in the bootstrap to look only among a subset of models, chosen on the basis of their information criterion score on the original data. This provides large computational gains with little apparent effect on inference. We also incorporate rigorous and economical ways of approaching issues of the existence of estimators when applying the method to sparse data tables.

Keywords Bias-corrected bootstrap · Computational economy · Human trafficking · Information criterion/criteria · Multiple systems estimation

1 Introduction

Multiple systems estimation has been used as an approach for quantifying hidden populations of many different kinds. It has been used specifically for human trafficking populations based on source data such as police records, non-governmental agencies, and outreach services. For an overall survey taking a view both of the range of applications and of methodology, see Bird and King (2018). For a recapitulation of a number of multiple systems estimation methods in policy contexts along with a survey of relevant data sets, see Silverman (2020). With administrative data sets becoming more publicly available over time, and data collection procedures reaching a wider and larger part of human trafficking populations, computationally stable and inexpensive multiple systems estimation procedures are desired for both researchers and practitioners.

Poisson loglinear regression (Cormack 1989) is one of the standard approaches. Recent developments include various approaches to model selection, and procedures tailored for sparse overlap in combinations of lists (Cruyff et al. 2017; Chan et al. 2021).

The loglinear method requires the choice of a model, specifying which terms are actually fitted to the data, and typically (Baillargeon and Rivest 2007; Silverman 2014) estimation is then conditional on the model selected for inference. Whatever approach is used in the frequentist paradigm, inference conditional on the selected model is likely to result in conservative standard errors and confidence intervals.

A bootstrap approach, for example as set out in Chan et al. (2021, Sect. 3.3), makes it possible in principle to account for the model choice step in the inference procedure. A stepwise model choice method may be sufficiently fast computationally to make the bootstrap feasible. However, a longer established approach to model choice (Baillargeon and Rivest 2007; Bales et al. 2015) is to choose among all possible models using the Bayes information criterion (BIC) and

✉ Bernard W. Silverman
bernard.silverman@stats.ox.ac.uk

Lax Chan
lax.chan@uniupo.it

Kyle Vincent
kyle.shane.vincent@gmail.com

1 Department of Statistics, University of Oxford, Oxford, U.K.

2 Dipartimento di Studi per l'Economia e l'Impresa, Università del Piemonte Orientale, Novara, Italy

3 Ottawa, Canada

Published online: 21 November 2023

the first aim of this paper is to extend the bootstrap methodology to this method. The focus on the BIC is intended as illustrative not prescriptive; a topic for future work would be a comparison with other model choice criteria in the bootstrap context.

If there are several lists, then the number of possible models is large and, as the BIC-based approach considers every possible model, the bootstrap approach leads to a computationally expensive and time-consuming procedure. In response to this limitation, we develop a modified bootstrap procedure that, for each bootstrap replication, only considers models that have high-ranking BIC scores on the original data. We also give some consideration to other notions of closeness to the best-fitting model. While we have only considered data sets which do not have concomitant information such as time dependence, the basic principles of our approach can be extended if such information is available and one wishes to choose between a large number of models. It is unfortunate that there are, to our knowledge, no publicly available data sets of that kind.

The current paper has been written from a frequentist viewpoint to encourage model selection to be taken into account when assessing estimation accuracy. There are several Bayesian approaches which, of their nature, avoid a focus on a single selected model. King and Brooks (2001) pioneered the use of a Markov Chain Monte Carlo approach to model fitting and averaging in this context. A comparison of a number of frequentist and Bayesian methods, including an MCMC approach, is given by Silverman (2020).

A second focus of this paper is on dealing rigorously and economically with issues that arise when some combinations of lists are not represented in the observed data. This sparseness of data is typical in the human trafficking context. This part of the work builds on the results of Fienberg and Rinaldo (2012a, b), which deserve greater attention in this area.

Detailed consideration is given to the UK Home Office data set on human trafficking victims (Home Office 2014), a data set on human trafficking in the New Orleans area (Bales et al. 2020), a data set on death occurrences in the Kosovo conflict from the late 1990s (Ball and Asher 2002), and a data set on sex-trafficked women in Korea (Ball et al. 2018). The results indicate that our modification of the bootstrap procedure provides significant computational savings without altering the inference substantively. The newly-developed bootstrap procedure is applied to multiple systems estimation data sets, but the broad approach has obvious utility and application to other contexts where a statistical procedure involves the choice between a large number of candidate models using some numerical criterion and a bootstrap approach is used to carry out inference taking into account the effect of model selection.

The paper is outlined as follows. Section 2 details the Poisson loglinear model, and sets out and develops the proposed bootstrap approach. Section 3 considers the construction and existence of estimators when the model is applied to sparse multiple systems estimation data sets. Section 4 presents results based on empirical applications. Section 5 explores further ideas for searching for optimal models within bootstrap resamples. Software to implement the methods of the paper and to reproduce the results is available in the most recent version of the SparseMSE R package (Chan et al. 2023).

2 The model and bootstrap

This section presents the Poisson loglinear model, sets out how a bootstrap approach can take account of model selection, and develops the idea of substantially reducing the computational burden by restricting the models considered in the bootstrap step.

2.1 The Poisson loglinear model

Suppose that there are t lists on which observed cases may fall. The set of lists on which a case is present is called its capture history ω, a subset of $\{1, \ldots, t\}$. Define the order of a capture history to be the order of this subset. Define the capture count N_ω to be the number of cases with capture history ω.

Let N^{obs} be the vector of the $2^t - 1$ observable capture counts $\{N_\omega : \omega \neq \emptyset\}$. Define $N_{\text{total}} = \sum_{\omega \neq \emptyset} N_\omega$, the number of cases actually observed. The count $N_\emptyset$, not included in N^{obs}, is the 'dark figure' of cases that do not appear on any list.

A model is defined by specifying a collection Θ of capture histories, always including the capture history $\emptyset$ and the t capture histories of order 1. The model has parameters α_θ indexed by the capture histories in Θ, and the capture counts are modelled as independent random variables with

$$N_\omega \sim \text{Poisson}(\mu_\omega); \quad \log \mu_\omega = \sum_{\theta \subseteq \omega, \theta \in \Theta} \alpha_\theta. \tag{1}$$

Note that both the capture counts N_ω and the model parameters α_θ are indexed by capture histories. Wherever possible we use ω as a suffix when the capture history refers to a count and θ when it indexes a parameter.

The parameters can be estimated by a generalized linear model approach, yielding parameter estimates $\hat{\alpha}_\theta$ and hence expected capture count estimates $\hat{\mu}_\omega$. The dark figure $N_\emptyset$ has expected value $\exp \alpha_\emptyset$ and so an estimate of the total population size M is $\exp \hat{\alpha}_\emptyset + \sum_{\omega \neq \emptyset} N_\omega$.

It is natural to restrict attention to hierarchical models defined to have the property that if θ is in Θ then so are all θ' that are subsets of θ. This can still leave a large number of

possible models. For example, for five lists, there are 6893 possible hierarchical models, excluding the non-identifiable model including all capture histories. In general, let $\mathcal{H}_{t,l}$ be the set of all hierarchical models for t lists allowing parameters/interaction effects of order up to l. If l is omitted it will be assumed that $l = t - 1$.

2.2 The Bayes information criterion

The Bayes information criterion or BIC (Schwarz 1978) is one of the approaches to model choice implemented by Baillargeon and Rivest (2007), and is one of the standard methods for multiple systems estimation (Cruyff et al. 2017; Chan et al. 2021). For any particular model Θ, the BIC value BIC_Θ is defined as

$$\text{BIC}_\Theta = |\Theta| \log N_{\text{total}} + 2 \sum_{\omega \neq \emptyset} (\hat{\mu}_\omega - N_\omega \log \hat{\mu}_\omega + \log N_\omega!). \quad (2)$$

To apply the BIC method, for t lists allowing for models of order up to l, the model in $\mathcal{H}_{t,l}$ with smallest BIC_Θ is chosen.

There is some ambiguity about the appropriate sample size to use in the definition. Baillargeon and Rivest (2007) consider the sample size to be the *case sample size* N_{total}, the total number of cases observed. On the other hand another possibility, regarding the table of capture counts as the data, is to use the *capture sample size* $2^t - 1$, the number of observable capture counts. There appears to be no strong argument of principle to determine which of these should be used. We use the case sample size because it has become standard in this field; however our general approach is equally applicable for the capture sample size, as well as for other scoring criteria for model choice.

2.3 The bootstrap

The bootstrap is a natural approach to take account of model selection when assessing the accuracy of estimation. The whole estimation procedure, including the choice of model, is carried out on each bootstrap replication. We focus attention on the bias-corrected and accelerated method set out by Efron and Tibshirani (1986). This method is second-order accurate and is invariant under transformation of the parameter space.

The multiple systems estimation context allows for various computational economies. The B bootstrap replications are obtained by simulating from a multinomial distribution with N_{total} trials and event probabilities proportional to the individual N_ω. For $i = 1, \ldots, B$, let $\hat{M}_i^{\text{boot}}$ be the estimate of the total population obtained from the ith bootstrap replication, and let $\hat{M}$ be the estimate from the original data.

The bias-corrected and accelerated method makes use of two control parameters, the bias-correction parameter and the acceleration parameter. The bias-correction parameter $\hat{z}_0$ is defined such that $\Phi(\hat{z}_0)$ is the proportion of the $\hat{M}_i^{\text{boot}}$ that are less than $\hat{M}$.

Estimating the acceleration parameter $\hat{a}$ involves a jackknife step, leaving out each of the original cases in turn and calculating an estimate of the population size from the resulting data. The following approach only requires the calculation of at most $2^t - 1$ estimates. Let Ω_1 be the collection of all capture histories for which $N_\omega > 0$. For each ω in Ω_1, let $\hat{M}_{(\omega)}$ be the estimate of the population size from the original sample but with N_ω replaced by $N_\omega - 1$. The number of estimates $\hat{M}_{(\omega)}$ is equal to $|\Omega_1| \leq 2^t - 1$. Once these have been calculated, the average of the jackknife estimates is given by

$$\hat{M}_{(\cdot)} = N_{\text{total}}^{-1} \sum_{\omega \in \Omega_1} N_\omega \hat{M}_{(\omega)}$$

and the acceleration factor by

$$\hat{a} = (1/6) S_2^{-3/2} S_3$$

where

$$S_k = \sum_{\omega \in \Omega_1} N_\omega (\hat{M}_{(\cdot)} - \hat{M}_{(\omega)})^k.$$

As set out byEfron and Tibshirani (1986), the bias-corrected and accelerated upper end-point of a one-sided β-confidence interval is the $\tilde{\beta}$ quantile of the $\{\hat{M}_i^{\text{boot}}\}$, where

$$\Phi^{-1}(\tilde{\beta}) = \hat{z}_0 + \{\hat{z}_0 + \Phi^{-1}(\beta)\}[1 - \hat{a}\{\hat{z}_0 + \Phi^{-1}(\beta)\}]^{-1}.$$

Note that, as is the case for the nonparametric bootstrap generally, the bootstrap replications are drawn with replacement from the original data, and therefore only the capture histories observed in the original data can be represented in the bootstrap samples. The possibility of drawing from a smoothed version of the original data when using the bootstrap was considered in some detail by Silverman and Young (1987). While in some circumstances smoothing can yield an improvement in the accuracy of estimation of a particular property of the underlying distribution, it can also possibly make matters worse. The details of any particular case are technical and somewhat obscure, and so it is safest to make the more conservative approach of sampling from the observed data only, though this is a possible topic for future research.

2.4 Restricting the models

Unfortunately the bootstrap computation can become prohibitive. If $t = 5$ and $l = 4$ then the model choice for

each data set requires the solution of 6893 generalized linear model problems. So if there are B bootstrap replications, the bootstrap process would involve fitting $6893B$ models, and therefore is computationally burdensome even with modern computing speeds.

To reduce the computation to manageable proportions, we approximate the bootstrap by restricting to a smaller number of models. Fix a value n_{top}, for example $n_{\text{top}} = 50$. Apply the Bayes information criterion approach to the original data, and let $\mathcal{H}_{t,l}[n_{\text{top}}]$ be the set of Θ in $\mathcal{H}_{t,l}$ giving the n_{top} smallest values of BIC_Θ. Then restrict the bootstrap and jackknife steps to the models in $\mathcal{H}_{t,l}[n_{\text{top}}]$. Each bootstrap and jackknife choice will only involve n_{top} models rather than the much larger number of all hierarchical models. The motivation is that even though a bootstrap or jackknife replication may result in a different choice of model from the original data, it is intuitively unlikely that a model that scored extremely badly on the original data would have the best score on the replication.

2.5 Varying the models considered

In assessing the effect of choice of n_{top} we calculate the bootstrap confidence intervals based on a range of values of n_{top}. This can be done without performing the full calculation for each value separately.

1. Fix the largest value $n_{\text{top}}^{\text{high}}$ to be considered.
2. For each bootstrap replication i and for each model j in $\mathcal{H}_{t,l}[n_{\text{top}}^{\text{high}}]$ find the population size estimate and the BIC. This gives $B \times n_{\text{top}}^{\text{high}}$ arrays of population size estimates $\hat{M}'_{ij}$ and of BIC values β_{ij}.
3. For each bootstrap replication i, find the indices of the record values of the vector $-\beta_{ij}$ with i fixed. Working recursively, let $j_0 = n_{\text{top}}^{\text{high}} + 1$ and, as long as $j_k > 1$, define $j_{k+1} = \arg\min_{j<j_k} \beta_{ij}$. This gives a decreasing sequence j_k. Finally construct a sequence of estimates $\hat{M}_{ij}$ by setting $\hat{M}_{ij} = \hat{M}'_{ij_k}$ for all $j \in [j_k, j_{k-1})$.
4. The array $\hat{M}_{ij}$ gives the bootstrap estimates for each value of n_{top} and for each bootstrap replication i. The same process can be carried out for the jackknife estimates. The bootstrap confidence intervals for every value of n_{top} can then be found.

The main computational burden is in the calculation of the individual $\hat{M}'_{ij}$ and β_{ij}, so this approach allows for the effect of all n_{top} up to a given value $n_{\text{top}}^{\text{high}}$ to be calculated for little extra effort than for the single value $n_{\text{top}}^{\text{high}}$. The maximum possible value for $n_{\text{top}}^{\text{high}}$ is $|\mathcal{H}_{t,l}|$, the total of all models considered. Increasing $n_{\text{top}}^{\text{high}}$ beyond this value will make no difference, so we can equally consider this to correspond to $n_{\text{top}}^{\text{high}} = \infty$, and this is the natural value to use for a general investigation of the approach.

3 Sparse data

3.1 Constructing estimators

In the context of modern slavery, particularly, it is common to observe data sets that are sparse in the sense that some of the N_ω are zero. For any capture history ω, define $N^*_\omega = \sum_{\psi \supseteq \omega} N_\psi$, the number of cases observed in the intersection of the lists in ω whether or not in conjunction with other lists. Define the parameter θ to be N^{obs}-empty if $N^*_\theta = 0$.

Chan et al. (2021, Section 2.4) set out an algorithm for sparse tables with models whose parameters have maximum order 2. The following algorithm is the natural generalization to N^{obs}-empty parameters of any order.

1. Initially define $\Omega^\dagger$ to be the set of all non-null capture histories and $\Theta^\dagger = \Theta$.
2. For each θ in Θ for which $N^*_\theta = 0$, record the maximum likelihood estimator of α_θ as $-\infty$ and remove α_θ from the set of parameters $\Theta^\dagger$ yet to be estimated.
3. For each such θ also remove from $\Omega^\dagger$ all ω for which $\omega \supseteq \theta$ (because $N^*_\theta = 0$ the corresponding N_ω 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 parameters in the model that are not estimated to be $-\infty$.

This approach allows the parameters to take values in the interval $[-\infty, \infty)$, extending the real line only at the negative end. The value $+\infty$ would not correspond to a finite parameter in the Poisson distributions of the observations, but $-\infty$ does make sense because these are parameters in a loglinear model, where a negative infinite parameter will correspond to a value of zero in the expected value, and hence the value, of the observed variable. Following Fienberg and Rinaldo (2012a, b) we use the term *extended maximum likelihood* for estimates where the value $-\infty$ is allowed. Far et al. (2021) provide additional information on determining parameter redundancy for a loglinear model along with a method for deriving the estimable parameters.

3.2 Existence of estimates

Fienberg and Rinaldo (2012b) show that the existence of the extended maximum likelihood estimate can be checked by

Table 1 Four possible observed data tables for a four list scenario

A	B	C	D	n_1	n_2	n_3	n_4
x				13	13	13	13
	x			16	16	16	16
		x		12	12	12	12
			x	11	11	11	11
x	x			3	3	3	0
	x		x	4	4	4	4
x	x	x		2	2	0	2
x		x	x	1	0	0	0
	x	x	x	1	0	0	0

the following linear programming problem. For $\theta \in \Theta^\dagger$ and $\omega \in \Omega^\dagger$ as obtained in the algorithm set out above, define $A_{\omega\theta} = 1$ if $\theta \subseteq \omega$ and 0 otherwise. Let ν be the vector of values N^*_θ for $\theta \in \Theta^\dagger$. Let s^* be the maximum value of s over all scalars s and all real vectors $x = (x_\omega, \omega \in \Omega^\dagger)$ satisfying the constraints

$$A^T x = \nu \quad \text{and} \quad x_\omega - s \geq 0 \text{ for all } \omega \in \Omega^\dagger. \tag{3}$$

A necessary and sufficient condition for an extended maximum likelihood estimate of the parameters to exist is that $s^* > 0$. We will then say that the Fienberg–Rinaldo criterion is satisfied. For any model where the criterion is not satisfied, we will define, by convention, $\text{BIC}_\Theta = \infty$, and so that model will not be chosen.

The criterion can only fail if there are zeroes in the data table. If there are no zeroes then setting $x_\omega = N_\omega$ for all ω will yield a feasible solution with $s = \min_{\omega \in \Omega^\dagger} N_\omega > 0$ and so the criterion will automatically be satisfied.

There is no obvious hierarchical relationship to the pattern of zeroes in the data that will lead to a model failing the criterion. Consider the example set out in Table 1. The data vector n_2 has two zeroes in positions where n_1 is non-zero, while n_3 and n_4 each have one more zero than n_2. For the model [123, 14] (A*B*C + A*D in standard R generalized linear model notation) the Fienberg–Rinaldo criterion is satisfied for n_1 and n_3 but not for n_2 and n_4. So while removing two cells from the support of n_1 to get n_2 means that the condition becomes violated, removing one more cell from the support may or may not restore the existence of the extended maximum likelihood estimate.

In conclusion, it seems advisable always to check the Fienberg–Rinaldo criterion if the data are sparse. We will discuss possible economies in this process in Sect. 3.3.

3.3 Economizing the check

Even after restricting to a smaller number of models, both the jackknife and the bootstrap steps involve evaluating the estimate for each particular model for a large number of possible data vectors. We show below that the existence of the extended maximum likelihood estimate depends only on the support of the observed capture counts and not on their actual values. This leads to the following approach, which allows for the Fienberg–Rinaldo checks to be carried out economically.

1. Remove from consideration all models which fail the check for the original data.
2. For the jackknife, check the condition for replications where the support is different from the original data, that is where $N_\omega = 1$.
3. For the bootstrap, find all the distinct supports among the bootstrap replications; this will typically be much lower than the number of bootstrap replications. For each support, construct a data vector which takes the value 1 on the support and 0 otherwise, and check the condition for that data vector. The result will hold for all bootstrap replications with that support.

Theorem The existence of the extended maximum likelihood estimate for a given model and data depends only on the support of the observed capture counts.

Proof: Suppose $N^{(1)}$ and $N^{(2)}$ are two arrays of observed capture counts with the same support. Given a model Θ, $\Theta^\dagger$ and $\Omega^\dagger$ only depend on the support of the data.

For the model Θ and the data $N^{(1)}$, suppose that the Fienberg–Rinaldo criterion is satisfied, with the vector x^1 satisfying (3) with $\min x^{(1)} = s^{(1)} > 0$. Now choose an integer k such that $kN^{(2)}_\omega \geq N^{(1)}_\omega$ for all ω, possible because the support of both vectors is the same. For ω in $\Omega^\dagger$, define $\tilde{x}_\omega = kN^{(2)}_\omega - N^{(1)}_\omega$. Define $\tilde{s} = \min_{\omega \in \Omega^\dagger} \tilde{x}_\omega \geq 0$. Then, for each θ in $\Theta^\dagger$, $(A^T \tilde{x})_\theta = kN^{(2)\star}_\theta - N^{(1)\star}_\theta = k\nu^{(2)}_\omega - \nu^{(1)}_\omega$.

Now define $x^{(2)} = k^{-1}(\tilde{x} + x^{(1)})$. Then $A^T x^{(2)} = \nu^{(2)}$, and for each $\omega \in \Omega^\dagger$ we have $x^{(2)}_\omega \geq k^{-1}(\tilde{s} + s^{(1)}) > 0$. Hence, for the model Θ and data $N^{(2)}$, the Fienberg–Rinaldo criterion is also satisfied. By applying the converse argument exchanging the roles of $N^{(1)}$ and $N^{(2)}$, we conclude that the criterion is satisfied for either both or neither of $N^{(1)}$ and $N^{(2)}$, completing the proof.

4 Empirical applications

We apply our bootstrap procedure to four real data sets. All analyses are, unless otherwise stated, based on 1000 bootstrap realizations.

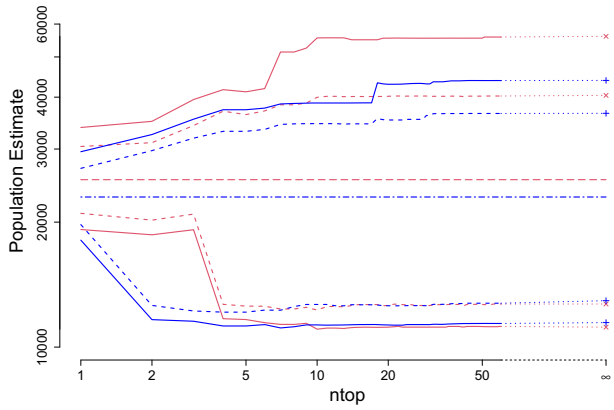

Fig. 1 Estimates and confidence intervals for the five-list UK data. Blue lines: maximum order 2; red lines: maximum order 4. Dot-dash horizontal lines: estimate; short dash lines: 80% confidence interval; solid lines: 95% confidence interval

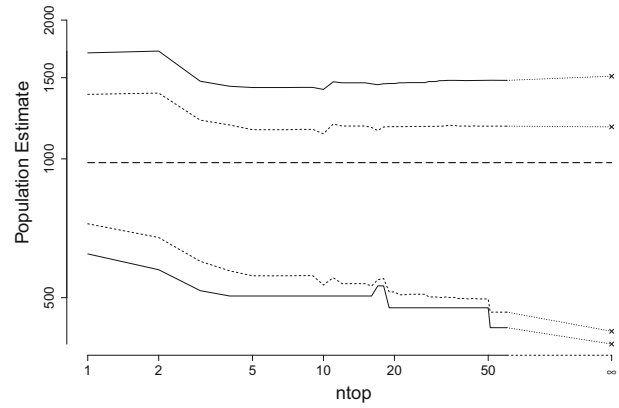

Fig. 2 Estimates for five-list New Orleans data. Dot-dash horizontal lines: estimate; short dash lines: 80% confidence interval; solid lines: 95% confidence interval. The results for models of maximum orders 2 and 4 are essentially identical

4.1 UK data

This five-list data set was studied in connection with the UK Modern Slavery Act 2015 (Home Office 2014) and discussed in detail in Bales et al. (2015). For other analyses and a fuller discussion, see for example Silverman (2020) and Chan et al. (2021). Figure 1 shows the estimates and confidence bands for the estimated abundance based on varying choices for maximum order and n_{top}. The value $n_{\text{top}} = \infty$ is used to denote the full calculation where all possible models are considered, 1024 models in the case of maximum order 2 and 6893 for maximum order 4. The interval for $n_{\text{top}} = 1$ is that obtained if we condition on the model chosen for the original data by the BIC criterion.

For both values of the maximum order, the results for $n_{\text{top}} = 60$ are virtually the same as those considering all models, while even n_{top} as small as 10 gives a result which is very close. Using $n_{\text{top}} = 10$ will reduce the computation time for the bootstrap stage by a factor of more than 100 (considerably more for maximum order 4.) Note also how much wider the confidence intervals are if allowance is made for the model choice. As one would expect, the bands are somewhat wider if models of order 3 and 4 are allowed.

4.2 New Orleans data

This data set (Bales et al. 2020) comprises 186 cases from New Orleans, distributed across 8 lists. A five-list version is produced by consolidating the four smallest lists into one. The results for maximum orders 2 and 4 are virtually identical, because the model with best BIC on the original data is of order 1, and all the other top models considered for $n_{\text{top}} \leq 60$ have order 2. All models with order 3 or 4 yield higher BIC values on the original data. There is a very small difference in the upper 97.5% confidence limit for $n_{\text{top}} = \infty$, which is 1547 for maximum order 4 and 1544 for maximum order 2. Values of n_{top} greater than 20 give confidence intervals that are slightly narrow but not dramatically different from $n_{\text{top}} = \infty$. An n_{top} over about 50 makes the difference even less.

4.3 Kosovo data

Probably the first application of multiple systems analysis in the human rights context was the estimation of the number of deaths in the Kosovo conflict in the late 1990s in connection with the trial of Slobodan Milosevich (Ball and Asher 2002; Ball et al. 2002). From a total of 4400 cases collated between four lists, the method arrived at an estimate of 10,356 Kosovar Albanian deaths. The original analysis narrowed down to models which fitted the original data table well enough to yield a chi-squared statistic with a p-value of at least 0.05, but discarded those which fitted too well, placing an upper cutoff at $p = 0.3$. Among the models remaining, the choice was made by minimizing the chi-squared statistic divided by the residual degrees of freedom. In fact this model also minimizes the BIC.

The bootstrap 95% confidence interval conditioning on this model is [9100, 12000], rounding to the nearest 100; this accords closely with the profile likelihood 95% confidence interval [9000, 12100] reported by Ball et al. (2002). It is instructive to construct confidence intervals taking account of the model choice, and these are displayed in Table 2. Even a value of n_{top} as small as 5 makes some allowance for the model choice; setting $n_{\text{top}} = 10$ gives the same confidence levels (within the rounding error) as if all the models were considered at the bootstrap stage, with more than a tenfold improvement in computational burden.

We also considered the possibility of applying the bootstrap to the original method of model choice. Unfortunately,

Table 2 Bootstrap confidence intervals for the Kosovo data, rounded to nearest 100

n_{top}	80% CI	95% CI
1	[9500, 11300]	[9100, 12000]
5	[8500, 11500]	[8500, 17000]
10	[7400, 12200]	[6900, 18000]
113 (∞)	[7400, 12200]	[6900, 18000]

Table 3 Bootstrap confidence intervals for various values of n_{top}, for the Korea data

n_{top}	80% CI	95% CI
1	[136, 198]	[131, 248]
2	[135, 286]	[130, 348]
6 (∞)	[135, 288]	[128, 349]

of the 1000 replications generated, 67 did not yield any models at all with chi-squared p-values in the specified range. With these replications excluded, the bootstrap 95% confidence interval is [7000, 15800]; however because of the exclusion of 67 replications there is no clear justification for this result.

4.4 Korean data

This data set was presented and analysed by Ball et al. (2018) in connection with the quantification of Korean women forced into sexual exploitation in a particular location during the Japanese occupation of Indonesia in World War II. There are three lists, labelled B, C and D. The total number of cases observed was 123, of which 12 occur on all three lists, 54 and 6 on the two-list intersections B∩C and B∩D, and 5, 5 and 41 on the single lists B, C and D respectively. The authors' original analysis used a Bayesian nonparametric latent-class model approach (Manrique-Vallier 2016, 2017) and provided a population estimate of 137, with a 95% credible interval of (124, 198).

Because there are three lists, there are eight possible models of order up to 2. Two of these, [12, 13] and [12, 13, 23], fail to satisfy the Fienberg–Rinaldo criterion on the original data and are removed from further consideration. The model [12, 23] has the lowest BIC score and yields a point estimate for the total population of 157.2, considerably higher than that produced by the method of Ball et al. (2018). The 80% and 95% confidence intervals for various values of n_{top}, rounded to the nearest integer and based on 1000 realizations, are as shown in the table. The confidence intervals are all shifted upwards relative to the credible interval found by Ball et al. (2018).

The result for $n_{\text{top}} = 1$, conditioning on the best fitting model, gives considerably narrower confidence intervals than those taking the model selection into account. However, restricting to the two top models gives results virtually identical with those obtained by considering all six models for each realization. There is no pressing computational need to restrict the number of models considered, but even in this case that number can be substantially reduced without materially affecting the result.

Table 4 For each data set and for each given value of n_{top}, the number of bootstrap replications (out of 1000) for which the minimum BIC value is obtained within the n_{top} best models on the BIC criterion for the original data

n_{top}	1	5	10	50	100
Kosovo	375	929	997	1000	1000
UK (maxorder 2)	216	583	739	930	954
UK (maxorder 4)	156	622	732	898	943
New Orl (maxorder 2)	357	622	698	915	958
New Orl (maxorder 4)	357	621	697	913	956

4.5 Further considerations

The bootstrap runs for $n_{\text{top}} = \infty$ facilitate a closer examination of the intuitive assertion that the BIC values from bootstrap replications are close to those from the original data. Let ρ_i be the Spearman rank correlation coefficient between the BIC values of the original data and the ith bootstrap replication, considering all models where both estimates exist. The average of the ρ_i is around 0.97 for the Korea, Kosovo and New Orleans data. For the UK data it is 0.86 or 0.88 for models of orders up to 2 and 4 respectively.

Another approach is to count the number of bootstrap realizations for which the global minimum of the BIC is achieved within the n_{top} best models on the original data. See Table 4. The best model for the original data is the best model for only 15% to 38% of the bootstrap replications.

5 Other possible approaches

5.1 Other measures of closeness

The discussion so far has been based on the intuitive idea that the best model for a bootstrap replication was likely to be one that scored well on the original data. In this section, we consider models that are close to optimal on the original data in a broader sense.

Define the distance function between two models $\delta(\Theta, \Theta^*)$ to be $|\Theta \triangle \Theta^*|$, the size of the symmetric difference between the two sets of parameters. Define Θ and Θ^* to be k-neighbours if $\delta(\Theta, \Theta^*) \leq k$.

Table 5 Bootstrap confidence intervals for various values of n_{top}, for the Kosovo data. The approach using the BIC values only is referred to as degree 1, while degree 2 uses the BIC ranks of degree 2, tie-breaking using the original BIC ranks

n_{top}	degree	95% CI
1	1,2	[9100, 12000]
5	1	[7900, 17400]
	2	[8000, 13500]
10	1	[6900, 18000]
	2	[8000, 17700]
113 (∞)	1,2	[6900, 18000]

For any given model Θ, define the BIC rank $r_1(\Theta)$ to be the rank of the BIC of the fit with model Θ to the original data, so that the best fitting model has $r_1(\Theta) = 1$. For integers $k \geq 2$, define $r_k(\Theta)$, the BIC rank of degree k, to be the smallest BIC rank of any $(k-1)$-neighbour Θ^* of Θ. The BIC ranks of all degrees up to any given K can be found recursively using the following algorithm.

1. Find the BIC ranks $r_1(\Theta)$ of all models from the estimation process using the original data.
2. For each Θ, let $\mathcal{N}(\Theta)$ be the set of all 1-neighbours Θ^* of Θ, including Θ itself.
3. Find the BIC ranks of increasing degree recursively, by finding

$$r_k(\Theta) = \min_{\Theta^* \in \mathcal{N}(\Theta)} r_{k-1}(\Theta^*) \text{ for } k = 2, \ldots, K.$$

The BIC ranks can be used in various ways to explore notions of closeness to the best model. As before, the basic principle is to narrow down the class of models considered, and then only to construct the bootstrap estimates on that smaller set of models. For example, order the models in order of their BIC rank of degree 2, splitting ties by reference to their original BIC rank, and restrict attention to the n_{top} models according to this ordering. We refer to this as the "degree 2" approach and the original method as "degree 1". Because both approaches will put the best model for the original data at rank 1, they will give identical results for $n_{\text{top}} = 1$, while for $n_{\text{top}} = \infty$ both approaches will also give identical results.

Table 5 gives results for the Kosovo data. For $n_{\text{top}} = 10$, the confidence intervals using only the BIC values are virtually identical with those obtained considering all 113 possible models; there is a very small difference in the upper limit which is subsumed in the rounding. However this approximation is not achieved by the degree 2 method.

See Fig. 3 for the UK and New Orleans data. For the UK data the degree 2 approach requires a larger value of n_{top} to obtain confidence intervals close to those for $n_{\text{top}} = \infty$, though once $n_{\text{top}} = 60$ there is little to choose. For the New Orleans data the two approaches give virtually the same results. Restricting to models of order 2 has virtually no effect for either data set.

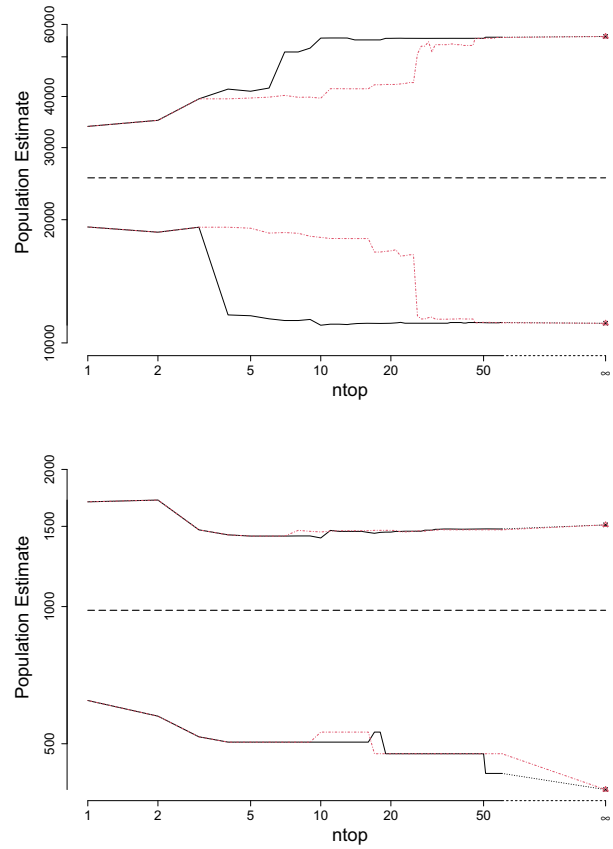

Fig. 3 Estimate and 95% confidence intervals for values of n_{top} up to 60, and $n_{\text{top}} = \infty$, allowing models up to order 4. Dashed line: estimate; black continuous: BIC values only used; red dotted: method using BIC ranks of degree 2. Top: five-list UK data; bottom: five-list New Orleans data

For another comparison, for each replication j, suppose that the model which minimizes the BIC over all models is at position $m_j^{(1)}$ for the degree 1 approach and $m_j^{(2)}$ for degree 2. For any realization j whose best BIC is observed for the best BIC model for the original data, $m_j^{(1)} = m_j^{(2)} = 1$. Comparative box plots omitting such realizations are presented in Fig. 4. The rankings of the individual optimum BIC models are clearly lower for the degree 1 than the degree 2 method, a conclusion confirmed by Wilcoxon paired signed-rank tests. The overall conclusion of this section is that considering neighbourhood information about models is an unwarranted complication.

5.2 Greedy search for BIC minimum

An exhaustive search for the BIC minimum involves a large number of models, and is infeasible for more than about five lists. Even with four or five lists, bootstrap approaches require approximations. Furthermore, if there is a desire to conduct

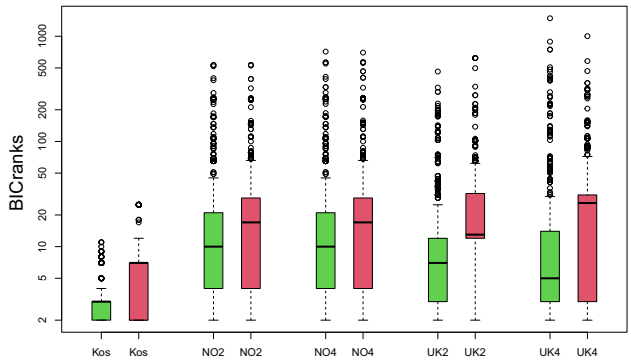

Fig. 4 Boxplots of ranks (log scale) on original data of optimal BIC over bootstrap replications, omitting replications where the ranks are of value 1. Green plots: ranks of BIC itself; red plots: ranking using BIC ranks of degree 2. Data sets/models, left to right: Kosovo; New Orleans with models of maximum order 2 and 4 respectively; UK maximum order 2 and 4 respectively

simulation studies then the computational burden becomes even more of an issue.

A possible alternative is a greedy search from one or more starting points to find at least a local minimum of the BIC or other criterion of interest. The obvious starting point is the order one model with no higher-list terms. At each stage all the neighbours of the current model are considered, obtained by either adding or removing a capture history from the parameter list, but only such that the hierarchical model property is preserved.

Because this method is quick to compute, it is practicable to repeat the estimation for each bootstrap replication. There is no longer any need to restrict to only five lists, so for the UK and New Orleans data results were also found for the full data (six and eight lists respectively). Results are presented in Table 6.

For the UK five-list data the approach gives a lower estimate and narrower confidence intervals than those obtained from the consideration of all models. To investigate further, we used the best minimum from five starting models of order 2, each containing five randomly chosen pairs of lists as parameters. The same starts were used for the jackknife and bootstrap replications. The point estimates and confidence intervals are virtually the same as if all models are considered. A relatively small number of random starts for the five-list data thus exposes the existence of the second estimate whose BIC is lower than the original. Silverman (2020) showed that, if a Bayesian approach is taken, the posterior for the UK five list data demonstrates an analogous bimodality.

For the UK six-list data, it is no longer feasible to consider all models. However as a check, downhill searches from random starts were also considered. Even using as many as fifty starting points never resulted in any estimate different from the one obtained from a single start from the null model. Although some downhill starts arrive at an estimate of around 21000, the BIC value is noticeably larger than that obtained by the search from the null model.

The conclusion is that if downhill searches are used, it may be worth investigating random starts with the original data as a check; if no secondary minimum of the BIC is found, then it seems safe to proceed with downhill searching only from the null model. As for the conclusion that should be drawn for the UK data, working with the six list data suggests a point estimate of around 12k with a confidence interval of 10k to 28k. The estimate and lower confidence limit are similar to those obtained by a different model selection approach in the original paper (Silverman 2014), though the upper limit is much higher. From a policy point of view, this is reassuring, because it is widely acknowledged that the original estimate obtained in Silverman (2014) is conservative.

6 Concluding remarks

Multiple systems estimation can lead to the need to choose between a very large class of models.The approaches we have

Table 6 Bootstrap 95% confidence intervals for various datasets, comparing the downhill search approach with the consideration of all models. Numbers correct to the nearest integer

	downhill search estimate	95% CI	all models estimate	95% CI
UK5 all models	12262	[9330, 32115]	25311	[11168, 56052]
UK5 maxorder 2	12262	[9332, 28954]	22991	[11456, 43894]
UK all models	12350	[9622, 27889]	six lists	
UK maxorder 2	12350	[9643, 25508]		
NewOrl5 all models	981	[397, 1564]	981	[397, 1511]
NewOrl5 maxorder 2	981	[397, 1564]	981	[397, 1505]
NewOrl all models	1110	[586, 1512]	eight lists	
NewOrl maxorder 2	1110	[586, 1512]		
Kosovo all lists	10357	[7133, 18681]	10357	[6947, 18010]
Kosovo maxorder 2	14342	[12525, 16477]	14342	[12451, 16273]
Korea	157	[128, 349]	157	[128, 349]

explored show that it is possible, using appropriate approximations, to use computer-intensive approaches such as the bootstrap to take account of the model choice in making inferences from the data. While we have concentrated on selecting a single model using BIC, our general approach is more widely applicable, for example if a different information or selection criterion is used, or if results are obtained by an ensemble of models weighted by reference to BIC or similar values. In general, inference conditional on the selected model or models can lead to optimistic statements about the uncertainty of estimates, and it is hoped that the work presented here will help to facilitate more robust assessments of estimation uncertainty.

Acknowledgements Lax Chan holds a career grant supported by the European Commission - FSE REACT-EU, PON Ricerca e Innovazione 2014–2020 and is a member of the Gruppo Nazionale per l'Analisi Matematica, la Probabilitá e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

Author Contributions BWS was the lead author. All authors reviewed and contributed to the main manuscript text. LC and BWS wrote the supporting computer package. All authors reviewed the package.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Baillargeon, S., Rivest, L.-P.: Rcapture: loglinear models for capture-recapture in R. J. Stat. Softw. **19**(5), 1–31 (2007)

Bales, K., Hesketh, O., Silverman, B.: Modern slavery in the UK: how many victims? Significance **12**(3), 16–21 (2015). https://doi.org/10.1111/j.1740-9713.2015.00824.x

Bales, K., Murphy, L.T., Silverman, B.W.: How many trafficked people are there in Greater New Orleans? Lessons in measurement. J. Hum. Traffick. **6**(4), 375–387 (2020). https://doi.org/10.1080/23322705.2019.1634936

Ball, P., Shin, E.H.-S., Yang, H.: There may have been 14 undocumented Korean "comfort women" in Palembang, Indonesia. Human Rights Data Analysis Group. Report available at https://hrdag.org/publications/there-may-have-been-14-undocumented-korean-comfort-women-in-palembang-indonesia/ (2018)

Ball, P., Asher, J.: Statistics and Slobodan: using data analysis and statistics in the war crimes trial of former president Milosevic. Chance **15**(4), 17–24 (2002). https://doi.org/10.1080/09332480.2002.10554820

Ball, P., Betts, W., Scheuren, F., Dudukovic, J., Asher, J.: Killings and Refugee Flow in Kosovo, March-June, 1999: A Report to the International Criminal Tribunal for the Former Yugoslavia. American Association for the Advancement of Science, Washington, DC (2002)

Bird, S.M., King, R.: Multiple systems estimation (or capture-recapture estimation) to inform public policy. Ann. Rev. Stat. Appl. **5**(1), 95–118 (2018). https://doi.org/10.1146/annurev-statistics-031017-100641

Chan, L., Silverman, B.W., Vincent, K.: SparseMSE: Multiple Systems Estimation for Sparse Capture Data in R. (2023). R package version 3.0.1. https://github.com/Laxchan/SparseMSE

Chan, L., Silverman, B.W., Vincent, K.: Multiple systems estimation for sparse capture data: Inferential challenges when there are non-overlapping lists. J. Am. Stat. Assoc. **116**(535), 1297–1306 (2021)

Cormack, R.M.: Log-linear models for capture-recapture. Biometrics **45**(2), 395–413 (1989)

Cruyff, M., Dijk, J., Heijden, P.G.M.: 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**(3), 41–49 (2017). https://doi.org/10.1080/09332480.2017.1383113

Efron, B., Tibshirani, R.: Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat. Sci. **1**, 54–75 (1986)

Far, S.S., Papathomas, M., King, R.: Parameter redundancy and the existence of maximum likelihood estimates in log-linear models. Stat. Sin. **31**(3), 1125–1143 (2021)

Fienberg, S.E., Rinaldo, A.: Maximum likelihood estimation in log-linear models. Ann. Statist. **40**(2), 996–1023 (2012a). https://doi.org/10.1214/12-AOS986

Fienberg, S.E., Rinaldo, A.: Maximum likelihood estimation in log-linear models: supplementary material. Available at http://www.stat.cmu.edu/~arinaldo/Fienberg_Rinaldo_Supplementary_Material.pdf (2012b)

Home Office: Modern Slavery Strategy. HM Government, London. Available at https://www.gov.uk/government/publications/modern-slavery-strategy (2014)

King, R., Brooks, S.P.: On the bayesian analysis of population size. Biometrika **88**(2), 317–336 (2001)

Manrique-Vallier, D.: LCMCR: Bayesian Non-Parametric Latent-Class Capture-Recapture. (2017). R package version 0.4.3

Manrique-Vallier, D.: Bayesian population size estimation using Dirichlet process mixtures. Biometrics **72**(4), 1246–1254 (2016). https://doi.org/10.1111/biom.12502

Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978). https://doi.org/10.1214/aos/1176344136

Silverman, B.W.: Modern Slavery: an application of Multiple Systems Estimation. Home Office, UK. Published in conjunction with the UK Government Modern Slavery Strategy; available at https://www.gov.uk/government/publications/modern-slavery-an-application-of-multiple-systems-estimation (2014)

Silverman, B.W.: Model fitting in Multiple Systems Analysis for the quantification of Modern Slavery: classical and Bayesian approaches. J. R. Stat. Soc. Ser. A **183**, 691–736 (2020)

Silverman, B.W., Young, G.A.: The bootstrap: to smooth or not to smooth? Biometrika **74**(3), 469–479 (1987)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.