does not happen. It would be interesting whether, based on this, one could build some graphical or numerical tool to be used as a general guideline for setting the thresholding prior parameters.

Ian Dryden, James Goulding and Simon Preston (*University of Nottingham*) **and Lax Chan**
(*The Open University, Milton Keynes*)
Professor Silverman's paper gives a fascinating overview of the challenges of estimating population sizes from sets of sparse overlapping lists of individuals. This is a difficult practical problem and requires careful modelling and close collaboration with stakeholders.

An alternative to thresholding is to consider a full Bayesian approach where a mixture prior is given for the interaction parameters in the model. We introduce latent variables $w_i \in \{0, 1\}$, $i = 1, \ldots, q$, for each of the q interaction parameters, with independent Gaussian mixture priors of two components 0 and 1 with distributions $N(0, \sigma_0^2)$ and $N(0, \sigma_1^2)$ respectively. The latent variable indicates component membership, with mixing proportions p_{mix} for component 0 and $1 - p_{mix}$ for component 1. It is straightforward and fast to simulate from the posterior by using MCMCpack (Martin *et al.*, 2011) for a given set of w_i, $i = 1, \ldots, q$, and then we use an outer loop of Metropolis–Hastings steps for switching w_i labels between 0 and 1 changing one at a time chosen at random.

In Fig. 9 we present some analysis where component 0 has a very small 'spike' variance $\sigma_0^2 = 1/1000$ to indicate that the parameter is effectively 0 with prior probability $p_{mix} = 0.9$ and $\sigma_1^2 = 1$. For these results the algorithm involves 1000 updates of the latent class variables and in between 1000 updates of all other parameters by using the existing Markov chain Monte Carlo (MCMC) algorithm with the appropriate prior precision matrix (which depends on the latent variables). We have thrown away the first 10% of each chain as burn-in and we have pooled 100 independent runs of the MCMC algorithm. The chains can remain in different modes where interaction parameters are not strongly identifiable, which leads to high or low estimates of the population size as seen in data sets UKdat_5 and Ned5. Such bimodality is also seen in Figs 2, 4, 5 and 6 in the paper. There is little in the data to distinguish between these parameters that lead to very different population estimates. It is difficult to see how this problem could be resolved without either more data or more overlap in the lists.

A final point is about the model assumption that individuals are homogeneous, which is clearly unrealistic. There are various models that take into account specific forms of heterogeneity (Baillargeon and Rivest, 2007). In simulation studies with heterogeneous individuals, we have found the pairwise interaction models of the paper quite effective, whereas the main-effects-only models clearly fail.

Peter G. M. van der Heijden and Maarten Cruyff (*University of Utrecht*)
We are not happy with the presentation of the Dutch data in the paper.

Professor Silverman reduces the number of lists from 6 to 5, taking together the I (Inspectorate) and O (residential treatment centres and shelters) as they have the lowest number of observations. However, such decisions should be driven by substantive grounds, e.g. that the organizations have similar objectives in making the lists and that the relationships between lists and background characteristics are similar. For the data constructed by Professor Silverman, it is unclear what they stand for. Below, we eliminate list K (Border Police), because many individuals on that list never entered the Netherlands because they were denied entrance at the border (see van Dijk *et al.* (2017)).

Contrary to what Professor Silverman states, covariates measuring background characteristics are usually available. For the United Nations Office on Drugs and Crime and the Walk Free Foundation we have produced multiple-system estimates for the Netherlands, the Republic of Ireland, Romania, Slovenia, Serbia and Slovakia, and for all these countries the data included covariates and multiple years.

Making use of covariates is essential. The instabilities that Professor Silverman observes are due to the presence or absence of certain two-factor interactions between lists. Inclusion of covariates, however, may replace list–list interactions by list–covariate interactions. This is because the covariates 'cause' list–list interactions when marginalizing over them. The International Working Group for Disease Monitoring and Forecasting (1995) have coined this *apparent dependence*.

To illustrate this, we first fitted all two-factor interaction models to the data with our choice of the five lists. Fig. 10 shows the relationship between the abundance estimates and the Akaike information criterion (AIC) and Bayesian information criterion (BIC) values of these models. Here the estimates lie in the range of 20000–70000: much smaller than the range in Fig. 4 of the paper.

All models in Fig. 10 with estimates in excess of 60000 have the interactions IZ, OP, OZ, PR and PZ in common. The models (IZ, OP, OZ, PR, PZ, RZ) and (IZ, OP, OZ, PR, PZ), as depicted by the black dots in Fig. 10, respectively have the lowest AIC and BIC values. Subsequently we have fitted these models
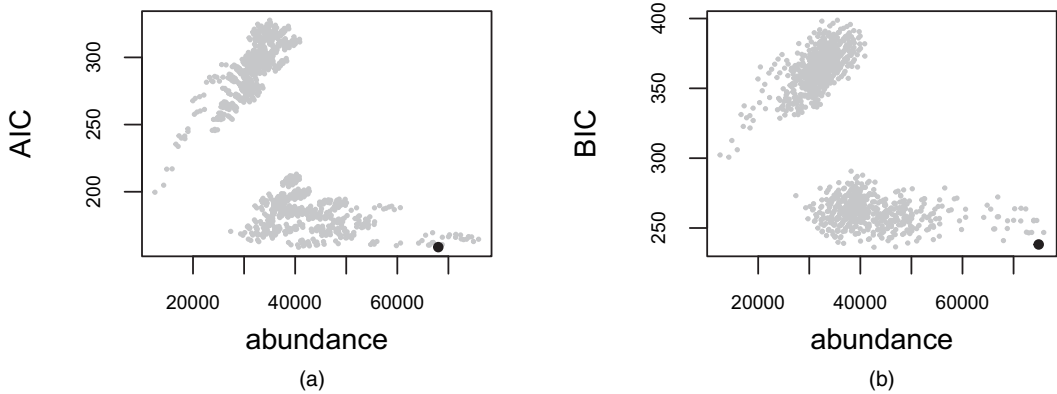
Fig. 10. Abundance estimates of all two-factor interaction models for the five-list Netherlands data based on (a) the AIC and (b) the BIC

with the main effects of the covariates sex, age, nationality and type of exploitation included, and then performed a stepwise model search. The AIC model yields an abundance estimate of 44000, with OP no longer in the model, and the BIC model yields an estimate of 42000, with IZ, OP, PR and PZ no longer in the model. The estimate produced in van Dijk *et al.* (2017) is also 42000.

James Jackson (*Lancaster University*)
The modern slavery data sets that are suitable for multiple-systems estimation (MSE) analysis are sparse. From Tables 1–3, the number of observable cells with 0-counts are, for the UK, 38/63 (60%), for the Netherlands, 39/63 (62%), and, for New Orleans, 236/255 (93%). Professor Silverman expects this to be the norm and has diligently developed methods to account for sparseness (see Chan *et al.* (2020)). Expecting sparseness is equivalent to expecting negative dependences between lists. Understanding why this is so is intrinsic for current methods to develop.

In the MSE model, for any list combination *A*, every individual in the population has the same probability of being in *A*. This may be unrealistic in a modern slavery environment, where several factors affect the visibility of a case; some poor victims may even have a probability of detection that is close to 0. The level of heterogeneity may also vary between lists, which would give rise to negative correlation. Agencies may have different abilities of reaching specific groups of the population. For example, a list deriving from the general public may be effective at identifying victims from car washes, but less effective at observing workers growing cannabis. Or an agency may be effective at identifying victims in one geographical area, but less so in another.

Moreover, undermatching would create a false impression of negative list dependences. Anonymized data make it difficult—or even impossible—to match individuals across lists, leading to one-list counts that are too high and multiple-list counts that are too low.

Neither the UK four-list data nor the Kosovo data (Table 4) contain any 0-counts, and estimates resulting from these data display impressive stability. Since MSE enables lists to be combined, the problem of sparse data could be relieved by combining lists in such a way as to minimize the number of 0-count cells. But, to begin to address the issue of sparseness, information about the list sources is essential. In the New Orleans data, even the names of the list sources are suppressed. To end with the 'fish in the lake' analogy, it needs to be checked that each list is catching from the same lake!

Antony Overstall (*University of Southampton*) **and Ruth King** (*University of Edinburgh*)
Multiple-systems estimation has a long history but has only recently been applied to modern slavery. Discussion of the multiple-systems estimation assumptions and their suitability to modern slavery would be useful, particularly as violations can have a significant effect on the population estimates (e.g. Overstall *et al.* (2014)).

A Bayesian threshold approach is used (applied to the ratio of the posterior mean to standard deviation) considering only two-way interactions. We raise two specific issues.

(a) Estimation of the total population size is dependent on a single model, and model estimates can vary substantially.