Supporting information for *Discussion on "The central role of the identifying assumption in population size estimation" by Serge Aleshin-Guendel, Mauricio Sadinle, and Jon Wakefield*

Daniel Manrique-Vallier

Contents

1 Description

This is a companion document to my discussion article regarding the manuscript "The Central Role of the Identifying Assumption in Population Size Estimation," by S. Aleshin-Guendel, M. Sadinle, and J. Wakefield (ASW). Here I revisit, discuss and, for the most part, dispute the results of their empirical investigation into the identifiability of Latent Class Models (LCMs) applied to the multi-list Capture-Recapture (CR) problem, as detailed in Section D ("Latent Class Models Simulations,") of the Web Supplement to their main manuscript, and discussed in their main document itself. Computer code to reproduce all the results from this document and the document itself is available at the Biometrics website on Wiley Online Library.

2 Introduction and Main Conclusions

The main objective of ASW's empirical investigation is to "demonstrat[e] the practical implications of [their] Theorem A.2." Said theorem establishes a sufficient condition for the statistical identifiability of the population size using LCMs for CR, and provides an explicit instance of failure when that condition is not satisfied—for a description of the notion of identifiability relevant to CR estimation, *conditional* identifiability, I refer readers to Section 2 of ASW's main article. The specific model they use for this evaluation is a (rather idiosyncratic) Bayesian *regularized* LCM for CR estimation (henceforth RLCMCR), originally proposed by Manrique-Vallier (2016) for CR estimation when the number of latent classes is unknown.

ASW's investigation is based on a collection of repeated-sampling experiments with CR data simulated from LCMs, which they feed to RLCMCRs with the objective of studying the procedure's performance at CR estimation. Each experiment revolves around a specific instance of LCM, crafted to resemble a realistic (thus "practical") situation. At the same time, ASW endowed each test LCM with structural features that violate the sufficient condition for identifiability from their theorem. Even though (with one exception) these models do not correspond to any of the identified examples of non-identifiability from their theorem, ASW *speculate* that they may nonetheless trigger estimation problems attributable to non-identifiability. The idea is that, if a test LCM constructed under these specifications leads to poor estimates, and this failure can be reasonably attributed to non-identifiability, and this non-identifiability can be claimed to be predicted by the theorem, then one could reasonably claim the experiment as an instance of the "practical [negative] implications of Theorem A.2."

Besides being intrinsically interesting as an inquiry into the properties of a specific statistical procedure, ASW's investigation also plays an important role in their argumentation in favor of their CR reform program. Indeed, ASW claim that their results "show [...] that [non-identifiability] is a practically relevant problem, as we have no guarantees for when estimates based on non-identified models are going to be accurate." Accordingly, ASW take their results as evidence of the urgency of barring the use of any CR method based on models whose conditional identifiability has not been established.

In this document I dispute most of the claims that ASW derive from their experiments. Specifically, I show that they have neither "demonstrated the practical implications of Theorem A.2", nor shown how non-identifiability is the impediment to CR estimation that they believe it to be. To be fair, even though my own, more *laissez-faire*, position is almost antithetical to ASW's, I do not claim to have refuted their *fundamental* concerns. However, I do show that the case they have presented is flawed enough to be dismissed. Additionally, from a more speculative perspective, I show that a closer look at some of their results can actually provide interesting elements to argue *against* some of their claims and, more broadly, their positions.

Among the problems that I identify and discuss in this document I can cite:

- Most experiments were designed in ways that make them plainly irrelevant to the claims ASW believe them to support.
- Most experiments (including those whose relevance can be reasonably claimed) involve a fair number of complex, interacting, factors which act as confounders. This alone is no argument to dismiss ASW's analysis (for one, ASW do acknowledge these limitations, even if they still attribute many of the results to the violation of Theorem A2 sufficient condition). However, some of these confounders have observable manifestations which, in conjunction with some understanding of RLCMCR models and

Markov chain Monte Carlo (MCMC) methods can help making sense of a fair share of the results. Without claiming to have demonstrated anything, in my re-analyses I take a closer look at some of these noteworthy results and attempt alternative (and, in my opinion, more sensible and interesting) interpretations.

- Related to the previous, one of these confounders is the use of RLCMCR models itself. RLCMCRs are regularized models which *a priori* penalize complexity—making their prior specification informative in that sense. Thus using RLCMCRs in the way ASW do (e.g. using the truncation of the stick-breaking series as if it were an *a priori* specification of the dimensionality of the model) makes it (unnecessarily) difficult to differentiate prior-induced bias from other effects. This is especially noticeable in those experiments where population size is not *trivially non-identifiable* (I define the term later) but the generating process violates theorem A2's sufficient condition. As I argue later (experiments 4b and 4c), I believe that this prior-induced bias led ASW to miss the most obvious explanation to the observed results.
- Model fitting has been performed using MCMC algorithms without performing the basic checks needed for avoiding demonstrably wrong computations, causing ASW to rely on spurious results. This, in my opinion, has caused ASW to miss some crucial insights, like the fact that truly non-identifiable LCMs lead to almost fixed-length posterior modal regions as the population size increases—which would have likely made them at least question their restrictive view on the utility of CR procedures under non-identifiability.

My conclusions about ASW's study notwithstanding, *absence of evidence is not evidence of absence*. ASW's failure at convincingly proving their point does not necessarily render their claims invalid. It might well be the case that, even if their theorem does not directly predict it, violation of the sufficient condition for identifiability does lead to the existence of yet-unknown but practically relevant non-identifiability regions in LCMs; maybe an open set in a plausible region of the parameter space. Therefore, empirical and theoretical work for identifying and characterizing non-identifiability regions (e.g. how big are these sets? where are they located? how big are the equivalence classes of undecidable models?), and actually understanding their *practical* consequences (e.g. are undecidable extrapolations different enough to worry about their multiplicity? are all undecidable models equally *reasonable*? is the size of the set of non-identifiable models big enough to care?) are still important pending tasks.

Finally, although I will not systematically argue for it here, my own position is that within some limits (some obvious like having *too many* more parameters than observations, others to be studied), non-identifiability in CR can be thought as more of a nuisance to manage, than an insurmountable obstacle to which we need to surrender. As I argue in my main text (See Section 3), whenever we know (or reasonably suspect) that a CR dataset was generated from a member of a family of distributions subject to *some form* of non-identifiability, attempting Bayesian CR inference with a model based on that same family is, if not the only, at least a sensible option. Point estimators *might* lack desirable properties, like consistency or the possibility of establishing convergence rates. However, posterior distributions will always be useful as summaries of how our knowledge about the parameters of interest improves after observing data, even if that improvement is less than what we could get from an alternative procedure based on an equally appropriate (but likely nonexistent) identifiable model. In Appendix 2 I present and discuss a case that exemplifies this point. I also do it in my re-analysis of some of ASW's experiments.

The rest of this document contains some background information and discussion on LCMs and RLCMCRs, a presentation and discussion of ASW's Theorem A.2, and my re-computation, re-evaluation, and impressions on ASW's experiments.

3 Background

3.1 Regularized Latent Class Models for CR

Latent Class Models (Lazarsfeld and Henry 1968) for *iid* multivariate binary data, $\mathbf{x}_i = (x_{i1}, \ldots, x_{iJ})$, are discrete mixtures of the form

$$\begin{aligned} x_{ij}|z_i &\sim \text{Bernoulli}(\lambda_{jz_i}), \text{ independently for } i = 1 \ldots N \text{ and } j = 1 \ldots J \\ z_i &\sim \text{Discrete}\left(\{1, \ldots, K\}, (\pi_1, \ldots, \pi_K)\right), \text{ iid for } i = 1 \ldots N. \end{aligned} \tag{1}$$

where the z_is are unobserved (hence latent).

A common interpretation of this model has it representing a heterogeneous population composed by individuals labeled $i = 1, ..., N$, each belonging to one of K homogeneous sub-populations, as indicated by the unobserved ("latent") individual-level class label z_i. Sub-population proportions are determined by the population-level vector of class probabilities $\boldsymbol{\pi}_K = (\pi_1, \ldots, \pi_K)$, which takes values within the $(K-1)$-dimensional probability simplex, Δ_{K-1}; i.e. is a probability mass function on the space of class labels. Lastly, at the sub-population level, it is assumed that individual item responses, x_{ij} are independent of one another; i.e. mixture components are independence models similar to main-effects log-linear models (Bishop, Fienberg, and Holland 1975).

Manrique-Vallier (2016) proposed the use of K^*-dimensional LCMs, paired with the priors

$$\begin{aligned} \lambda_{jk} &\sim \text{Beta}(a, b) \text{ iid for } k = 1 \ldots K^* \text{ and } j = 1 \ldots J \\ \boldsymbol{\pi}_{K^*} = (\pi_1, \ldots, \pi_{K^*}) &\sim \text{SB}_{K^*}(\alpha) \\ \alpha &\sim \text{Gamma}(d, e) \\ p(N) &\propto 1/N, \end{aligned} \tag{2}$$

for J-list Capture-Recapture (CR) estimation of N (*Regularized Latent Class Model CR*, RLCMCR). Here N represents the unknown population size, vectors $\mathbf{x}_i$ represent the capture history of each individual $i \in \{1, ..., N\}$ ($x_{ij} = 1$ for individual i captured by list j, $x_{ij} = 0$ otherwise), and $\text{SB}_{K^*}(\alpha)$ stands for the K^*-dimensional finite stick-breaking process with concentration parameter α, as described by Ishwaran and James (2001). CR data consists of the all vectors $\mathbf{x}_i$ such that at least one of their J coordinates is $x_{ij} = 1$; i.e. have been observed at least in one list.

A key feature of this specification is the regularizing effect induced by the prior on $\boldsymbol{\pi}_{K^*}$. This distribution, $\text{SB}_{K^*}(\alpha)$, was originally introduced as an approximation by series truncation to the stick-breaking representation of the Dirichlet process (Sethuraman 1994). Like its infinite-dimensional cousin, realizations from SB_{K^*} are K^*-dimensional discrete probability vectors where the bulk of the probability mass tends to concentrate within the first few coordinates. Thus, as a prior distribution for the vector of LCM sub-population proportions, this specification *a priori* favors sparse mixtures, where the first few sub-populations, as indexed by $k = 1, ..., K^*$, account for most of the population, while the rest become negligible in size. This behavior allows fitting RLCMCRs to data without knowing in advance the true number of sub-populations, $K \ll K^*$, as is almost always the case in practice. Indeed, posterior Bayesian inference using this specification will tend to keep the number of mixture components of non-negligible size as low as possible, while maintaining a good fit to the observed data. Therefore, RLCMCR is a method that estimates population sizes essentially by finding a sparse mixture of independence models that can account for the observed data and extrapolating from there.

3.1.1 Discussion: RLCMCRs in ASW's investigation

RLCMCRs are a poor choice for the type of investigation ASW are interested in. To understand why, we need to note two things. First, for this type of experiments to make sense, the data generating distribution must be member of the same family of distributions used for estimation. In ASW's Theorem A.2 this family consists of the LCM distributions where both the number of lists, J, and the number of latent classes, K, are fixed and known. Second, when correctly applied, the parameter K^* in RLCMCRs must be such that $K \ll K^*$. Therefore, the data generation process lies on the boundary of the parameter space of the family

that underlies RLCMCR—or, in other words, the generating distribution does not belong to the family that underlies RLCMCR.

The distinction between the parameter K^* of an RLCMCR, and the parameter K of an LCM is critical. RLCMCRs are indeed K^*-dimensional LCMs, but just *formally*. Functionally, however, the parameter K^* of an RLCMCR is just a computational device employed to approximate an infinite-dimensional prior process (stick breaking) using a finite-dimensional one (SB_{K^*}), under the assumption that $K \ll K^*$. In order to appropriately fit an RLCMCR to data, we need to let K^* to be as large as computational resources allow, and to rely on the regularizing properties of the $SB_{K^*}(\alpha)$ process to shrink the *effective dimensionality* of the fitted model to whatever is needed to account for the observed data. This shrinkage is not to zero, so the dimension of the generating model and the fitted model will be always different. As a consequence of this design, RLCMCRs do not provide a mechanism for *a priori* specifying perfect knowledge about the dimension of the generating process. Thus, if for any reason we actually knew the data generating process' dimension K, RLCMCRs would just not be an appropriate choice, as they do not offer any reasonable way of incorporating that knowledge.

ASW misuse RLCMCRs by treating the computational approximation parameter K^* as if it were an *a priori* specification of K. As explained above, this specification runs against RLCMCR's design and intended usage. And, while doing so *formally* forces RLCMCRs to be LCMs with K latent classes, *functionally* it results in prior specifications with non obvious properties—and certainly not uniform over the K-simplex—which have complex and difficult to assess effects. These effects might then interact with (or just get superposed to) the effects of the structures that ASW really wanted to understand. The final result is a poor experiment, where potential causes of noteworthy observable results are difficult or even impossible to tell apart.

In my re-computation of ASW's experiments I have tried to remove some of these unnecessary confounders by setting $K^* = 10$, which being larger than the largest experimental $K = 4$, should work fine in all cases (even if it leads to "non-identifiable" RLCMCRs). This setup allows me to anticipate and make sense of some results that are easily attributable to RLCMCR's properties, without needing to speculatively attribute them to hypothesized yet-to-be-discovered consequences of violating ASW's sufficient condition for identifiability.

I finally note that RLCMCRs would still be a poor choice for this investigation, even if they were correctly applied. As I have already discussed, RLCMCRs have a complex, non-neutral structure aimed at learning the dimensionality of the generating process from the data, without having to specify it *a priori*. If our objective is to spot observable results that can be attributed to non-identifiability (or any other cause) by setting controlled conditions that might trigger them, the complex structure of RLCMCRs will act as an experimental confounder by both interacting with the experimental conditions and by generating extraneous effects. An example of this are experiments 4b and 4c, where ASW's sufficient condition is violated, and results show a noticeable bias for small N, which tends to disappear as N increases. This is a case in which we cannot really say if ASW have finally found their instance of non-trivial non-identifiability with relevant observable consequences, or if the results are just a manifestation of prior-induced bias, as I suspect. These problems could be reduced by using a simpler model, that only retains the fundamental aspects of an LCM, thus reducing the number of moving parts of the experiment. I believe that using an LCM with a fixed K and where the prior for $\boldsymbol{\pi}_K$ is uniform over Δ_K could be a reasonable starting point.

3.2 Identifiability and ASW's Theorem A.2

Consider an LCM family indexed by the $K(J+1)-1$ dimensional parameter

$$\theta = (\boldsymbol{\pi}_K, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K)\,,$$

where $\boldsymbol{\pi}_K = (\pi_1, ..., \pi_K)$, and $\boldsymbol{\lambda}_k = (\lambda_{1k}, ..., \lambda_{Jk})$, as described in (1). As ASW note, the rather obvious condition on $\boldsymbol{\theta}$,

$$2^J - 1 < K(J+1)$$

is sufficient for this family to be non-identifiable. Models with these characteristics can be problematic if not paired with further restrictions or informative prior distributions. These models, however, are trivial instances of non-identifiability (just more parameters than observations), and therefore not all that interesting. Let us call them ***trivially non-identifiable*** models.

ASW's Theorem A.2 establishes two much more interesting results specifically regarding the identifiability of the population size in CR using LCMs. These results, in an informal presentation, are:

1. A **sufficient** condition for identifiability: Let $\Theta = \{\boldsymbol{\theta} : \boldsymbol{\pi}_K \in \Delta_{K-1}, \lambda_{jk} \in (0,1) \text{ for } j = 1...J, k = 1...K\}$, then
$$J \geq 2K \implies \text{Population size is identifiable for any } \theta \in \Theta.$$

2. An **explicit instance of non-identifiability** when the above sufficient condition fails: Let Θ, defined as in the previous, be the parameter space of an LCM with $J < 2K$. Then there exist two disjoint one-dimensional sets $A, B \subset \Theta$ and a one-to-one mapping $f : A \leftrightarrow B$ such that both $\theta \in A$ and $\theta' = f(\theta)$ index two different LCMs (thus imply two different expected population sizes), but imply the same observable CR distribution.

3.2.1 Discussion: Identifiability, Theorem A.2's scope, and ASW's Empirical Investigation

While the first part of ASW's theorem settles the question of identifiability when $J \geq 2K$ in a positive sense—and therefore is an important result for statistical practice—the practical relevance of part 2 is less clear. ASW's explicit construction of the set A and its counterpart, $B = f(A)$, is by ASW's own admission, rather artificial (e.g. it assumes that all lists have exactly the same probability of capturing members of the same latent class), and small (with measure zero under any reasonable prior distribution over Θ). Therefore, as ASW also acknowledge, it is unlikely that the theoretical problems that these sets might create will have any *practical* relevance.

These acknowledgements notwithstanding, ASW present part 2 of their theorem as a proof of non-identifiability for LCMs used for CR when $J < 2K$. This characterization is, of course, trivially correct when we consider the parameter space to be the whole set Θ, which makes $(A \cup B) \subset \Theta$ a counterexample. This is, however, not the only way of interpreting these results. Indeed, parameter spaces, just like any other part of the definition of a model, can be set up to be whatever we deem *useful* for a purpose, as long as the resulting model is logically consistent—as an example, note that statisticians routinely exclude the boundary of compact parameter spaces in theoretical investigations. Therefore, instead of considering these results to be the end of the line, as ASW do, we could use them to refine parameter spaces to avoid potential problems by, for example, defining them to be such that $\Theta^* \subseteq \Theta \setminus (A \cup B)$. Considering that $A \cup B$ is a 1-dimensional manifold embedded into a $(JK + K - 1)$-dimensional space, this re-arrangement is unlikely to have any practical effect.

Therefore, instead of taking $J < 2K$ to be a sufficient condition for non-identifiability in Θ, a more useful statement could take the form of a condition that the parameter space of an LCM, say Θ^*, would need to satisfy in order for the population size to be globally identifiable when $J < 2K$:

2. (alternative) A **necessary** condition for conditional identifiability when the sufficient condition fails: Let Θ^* be the parameter space of an LCM and assume $J < 2K$. Then
$$\text{Population size is identifiable} \implies (A \cup B) \not\subset \Theta^*,$$
for A and B defined according to ASW's construction.

Formalities aside, none of these considerations addresses ASW's true concern: given that the violation of the necessary condition $J \geq 2K$ implies the existence of the sets A and B, *it might also be the case* that there exist not-yet-discovered regions of non-identifiability in Θ (or Θ^*) which, unlike $A \cup B$, *might* be actually problematic in practice. These are the "practical implications of Theorem A.2" that ASW's claim to have illustrated with their investigation. Thus, consistent with their strong beliefs regarding the insurmountability of the negative effects of non-identifiability in statistical inference, ASW adopt the rather intransigent position of recommending to avoid LCM-based CR methods whenever one suspects that the generating process is such that $J < 2K$.

I do not share ASW's pessimism. While I do believe that ASW's investigation has failed at showing the dangers they anticipated—i.e. they have not shown that $J < 2K$ on its own leads to practical problems consistent with non-identifiability—and should not be regarded as evidence of anything, my real reasons are deeper. As I have argued in my main article and in the introduction to this document (and will continue

arguing throughout the rest of this supplement), I am convinced that, even if ASW were correct about the *existence* of these sets, their practical consequences would be more something to manage, than to avoid—more annoyances, than real dangers. These annoyances would certainly need to be understood and taken seriously, but their mere existence need not result in the uncritical avoidance of LCMs nor any other CR method based on models might exhibit some form of non-identifiability. Understanding this is specially important in cases in which we have reasons to believe that the actual data generation process belongs to a non-identifiable family. In such cases, posterior distributions from "non-identifiable models" could be the only sensible inference tool available to us.

4 Another look at ASW examples and concerns

ASW's empirical exploration comprises a series of repeated-sampling experiments designed to showcase what they believe are the "practical implications of Theorem A.2". Each experiment is constructed around a specific LCM, purposefully built to illustrate some aspect of ASW's claims. From each of these models they generate three groups of 200 synthetic CR datasets, each corresponding to a different population size of $N = 2000, 10^4, 10^5$, to which they then fit RLCMCR models, with K^* set to be equal to the generating process' number of latent classes, K. They finally combine the results to form empirical estimates of repeated-sampling properties (like bias or credible-interval coverage) of the posterior distribution of a parameter relevant to CR estimation. The parameter in questions is

$$p_0 = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} (1 - \lambda_{jk}),$$

which is the probability of not being captured by any of the J lists.

As I discuss in the introduction, my evaluation of ASW's investigation is that it suffers from important flaws in its design, execution, and interpretation of results, and that these flaws run deep enough to render it inapplicable to ASW's purposes. In this section and the appendices I re-examine ASW's experiments and offer evidence to back my criticism. I discuss each experiment design and suitability (or not) to its intended purpose. For this, after noting that a fair share of ASW's computations are inadequate, I re-do all of them. Finally, I discuss the results, criticize ASW's interpretations when granted, and offer my own interpretations and observations. The rest of this section contains my impressions and interpretations. I offer detailed experimental results in Appendix 1.

4.1 Two-Sample Examples (examples 1 and 2)

The first two examples in ASW's exploration are based on simulated data from LCMs with $J = 2$ lists and $K = 2$ latent classes. ASW declare these as non-identifiable because of the violation of their $J \geq 2K$ *sufficient* condition for identifiability. However, these are actually *trivially non identifiable* models, as they also violate $2^J - 1 \geq K(J+1)$—i.e. they just have too many parameters to estimate them all—which is a much stronger condition to violate. This means that any observable anomaly consistent with non-identifiability that we might want to *speculatively* link to Theorem A.2 (through the violation of $J \geq 2K$), could be more convincingly explained as a consequence of the *known-to-be-there* trivial non-identifiability. In other words, ***these examples are irrelevant*** to the efforts of showing the "practical implications of Theorem A.2."

The inadequacy of these examples for showing anything beyond the obvious consequences of trivial non-identifiability should not come as a surprise. After all, $J = 2$ can only provide three observable cells, which is just enough for fitting two parameters in addition to the population size—incidentally, this is also the reason why no serious practitioner would ever consider fitting models more complex than simple independence to two-list CR data.

To help understanding just how unreasonable it is to expect any sensible LCM inference from data with these characteristics (CR data generated from a trivially non identifiable mode with two lists and two latent classes), it might help to note that even if we could specify the number of latent classes in advance (which avoids RLCMCR-specific regularization effects), and ***use the whole population as data*** (therefore bypassing

all CR-specific problems), we would still be unable to estimate the model's parameters. The code below illustrates this. There I use the R package *BayesLCA* to simulate a population of $N = 100,000$ units from ASW's Example 1. I then feed this dataset *in its entirety* to an LCM fitting procedure supplied with the correct number of latent classes $K = 2$. I include some observations as inline comments within the code.

```
library(BayesLCA)

#For replicability. Could be changed if anyone wanted to.
set.seed(123)

# Example 1 from ASW
lambda <- matrix(c(0.2475, 0.7425, 0.2475, 0.7425), ncol = 2)
pi <- c(0.5, 0.5)

# Generate a sample with N=100,000
d <- rlca(100000, itemprob = lambda, classprob = pi)

# Try to fit the *complete data* (no CR) to the *actual* sampling distribution
# (Note the warning generated by the fitting procedure!!!)
x <- blca(d, G = 2, verbose = FALSE)
## Warning in blca.em(X, G, ...): Model may be improperly specified. Maximum number
## of classes that should be run is 1 .

# Let's see what we've got (spoiler: not even close.)
x$itemprob
##           [,1]     [,2]
## [1,] 0.6175536 0.617591
## [2,] 0.0000000 0.000000

x$classprob
## [1] 0.8012746 0.1987254
```

No surprises here.

The point is that ASW's Examples 1 and 2 do not illustrate anything specific to either Theorem A.2, RLCMCR, CR, or even LCMs. They are not instances of "the practical implications of theorem A.2." They are just instances of statistics' general inability of extracting information from data which do not contain it.

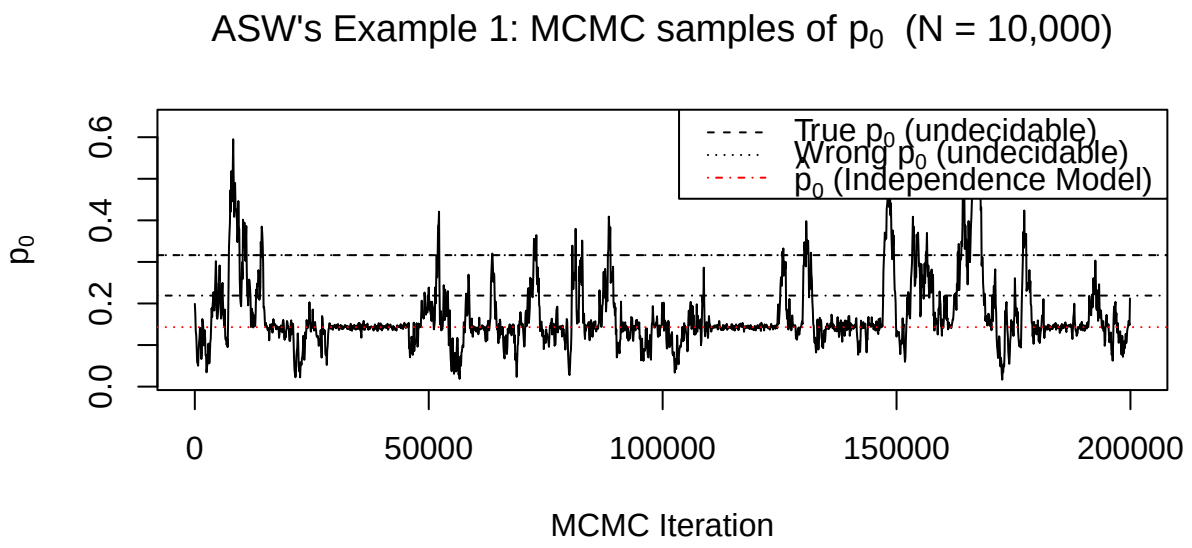
4.1.1 Some lessons from Examples 1 and 2

Inadequacy for their intended purpose notwithstanding, Examples 1 and 2 can still illustrate some interesting facts.

1) RLCMCRs are *regularized* mixtures of *independence* models As discussed above, RLCMCRs are essentially regularized mixtures of independence models. When CR data originate from just $J = 2$ lists, independence models are saturated—so they will always result in a perfect fit. This means that when RLCMCRs are applied to two-list data, fitting will always strongly favor configurations equivalent of one-component mixtures. Therefore RLCMCR's behavior will be dominated by the equivalent of an independence model.

This is exactly what we observe in ASW's examples 1 and 2. Take Example 1. This simulation was constructed from one of the problematic cases identified in part 2 of Theorem A.2; i.e. an LCM model with $K = 2$ latent classes for 2-list data, which induces the same observed data distribution as a different LCM of the same dimensions ($J = 2, K = 2$), while at the same time implying a different probability of no-capture, p_0. This sophisticated structure, however, turns up to be irrelevant in practice: RLCMCRs will, predictably, behave as an independence model regardless. This can be observed in the following trace plot of MCMC samples

from the posterior distribution of p_0, generated using the R package *LCMCR*. Here I used the same MCMC parameters as ASW (200,000 samples after a 50,000 sample burn-in period), and the same computational parameter $K^* = 2$ for RLCMCR fitting (although, unsurprisingly results are similar for larger K^*s).

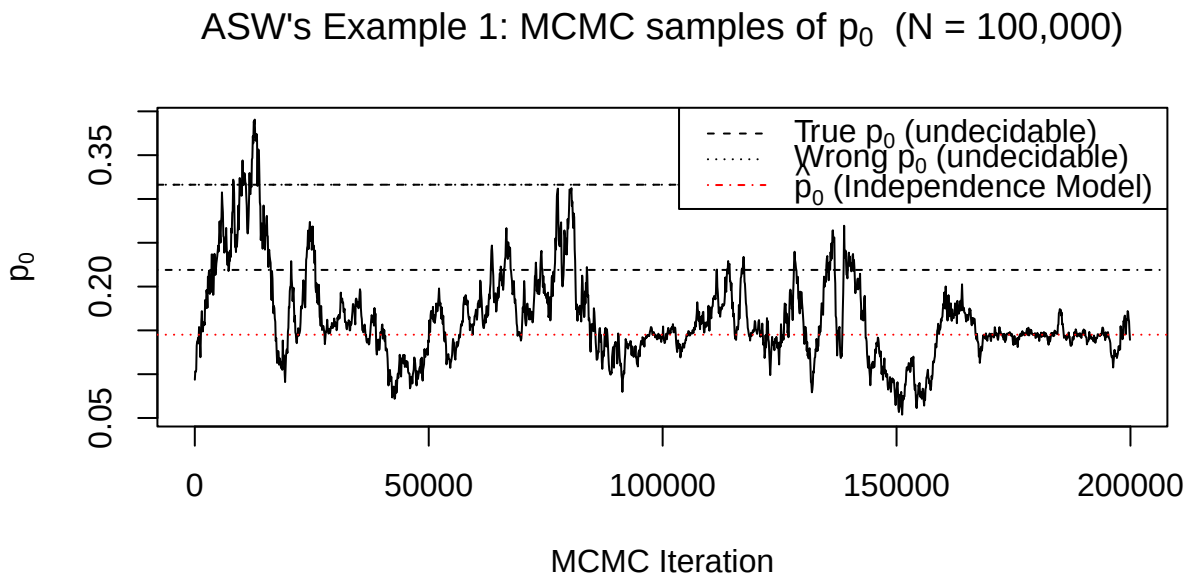


Leaving aside the fact that this is an obviously poorly mixed chain (which helps explaining some strange results obtained by ASW, especially for large N), we see that samples are *unsurprisingly* mostly concentrated around the Pertersen-type independence estimate $\hat{p}_0$ of p_0, indicated with a red line.

Just as interesting, we see that even though most posterior samples come from the immediate vicinity of $\hat{p}_0$, the chain often takes excursions to much farther away regions, which include both the true value and the undecidable-from-data value predicted by Theorem A.2. This occurs because while RLCMCR does *favor* simpler representations, it still puts prior positive mass on more complex mixtures. The specific behavior, however, is mostly dominated by the prior specification. The current implementation of *LCMCR* has subpopulation-level prior distributions (distribution of λs in (2)) hard-coded a to conservative specification (i.e. favor low class-level p_0s), when the number of lists is reasonably large (e.g. $J = 4$ implies *a priori* $E[p_0] \approx 0.0625$, and $\Pr[p_0 < 0.1] \approx 80\%$, while $J = 8$ implies $E[p_0] \approx 0.00391$, and $\Pr[p_0 < 0.0038] \approx 80\%$). This means that sub-populations where prior specification dominates over data will, by default, only contribute a negligible amount to the overall p_0. However, when the number of lists is too small, like in this case, this prior specification is more diffuse and centers around larger p_0s (e.g. $J = 2$, these prior induce numbers of lists (e.g. $J = 2$ implies *a priori* $E[p_0] \approx 0.25$, and $\Pr[p_0 < 0.43] \approx 80\%$). Therefore, this part of the specification (and maybe and to a lesser degree, the proximity to the true and the indistinguishable-from-truth values) can explain these deviations from the *predictable* RLCMCR output.

2) Posterior distributions can be complex, and MCMC output requires attention Posterior densities of mixture models are notorious for often being highly multimodal. RLCMCRs are no exception. Additionally, non-identifiability can result in posterior distributions having several equally important modal regions. Therefore, it is important to pay attention to the general shape of the posterior distribution, and not only to a few summaries like means or quantiles.

This is specially important when using estimation algorithms based on MCMC sampling. As we have seen, MCMC samples from RLCMCR posterior distributions can sometimes take very long runs to mix appropriately. This is even more important with large sample sizes, where narrower modal regions can sometimes "trap" the chain for long periods, and data-augmentation steps result in highly auto-correlated chains. Take, for instance, the following run corresponding to ASW's Example 1 with a population size of $N = 100,000$, and the same MCMC parameters used by the authors.



We see that, as expected, the mixing of this chain is worse than in the previous example, which was computed from data generated from the same LCM model but with a smaller population of $N = 10,000$. I also note that this computer output, which is unsuitable for MCMC estimation, was actually used by ASW (among several other equally faulty chains) as the basis of their analyses.

In my own experience, the MCMC algorithm implemented in the *LCMCR* package can sometimes require very long runs, in the order of several million iterations, to allow a reasonable assessment of the exploration of the support of the posterior distribution.

These problems are, of course, not exclusive of *LCMCR*, but are a general limitation of MCMC methods. As it is well known, it is usually impossible in practice to guarantee that an MCMC run of any length will generate a collection of samples representative of a posterior distribution (although in many cases, like the ones shown here, it is possible to determine that a chain *is not* adequate). So here I just echo Charles Geyer recommendation regarding MCMC practice: "[T]he least one can do is to make an overnight run" (Geyer 2011).

4.2 The other examples (3a, 3b, 4a, 4b and 4c)

The rest of ASW's examples are mixed bag. In total ASW present 5 additional examples. From them, two are trivially non-identifiable (examples 3a and 4a), one is fully identifiable as proven by Theorem A.2 (example 3b), and two are not trivially non-identifiable but violate the condition $J \geq 2K$ so are described by ASW as "not conditionally identifiable" (examples 4b and 4c). Therefore only three out of the 7 experiments are actually relevant to ASW's claim of illustrating the *"practical implications of Theorem A.2"*: 3b (for part 1), and 4b and 4c (for part 2).

Heeding the lessons from the previous examples, I have re-computed all of these experiments by extending the individual MCMC runs, from ASW's clearly insufficient 200,000 iterations per run, to 10 million. To save space, I have also sub-sampled ("thinned") this output by retaining one every one thousand iterations, for a total of 10,000 samples per run. Finally, I have set RLCMCR's approximation parameter $K^* = 10$, in order to allow the regularization machinery to work as intended.

In the following section I detail my impressions regarding the results of these five experiments. The detailed results can be found in Appendix 1, at the end of this document.

4.2.1 Comments on the results of the experiments

Fully identified model (3b) This is the best behaved case in the whole lot. It illustrates well that under these conditions, as predicted by Theorem A.2 (part 1), RLCMCR exhibits the basic properties that one

would expect from a good statistical procedure based on the true generating process under identifiability: approximately unbiased point estimates, credible intervals have approximate nominal coverage and seem to shrink for increasing sample sizes.

I would like to emphasize that the RLCMCR procedure was set up with $K^* = 10$, making the model not only nominally "non-identifiable," but according to ASW usage, a different parametric family from that of the generating model. Nevertheless, the procedure worked as intended, finding a structure within the RLCMCR space which was roughly equivalent to that of the generating model.

Trivially non-identifiable models (3a and 4a) As I already argued at the beginning of Section 4.1, trivially non-identifiable models cannot serve as illustrations of the consequences of ASW's Theorem A.2. Nonetheless, there are many other lessons regarding CR estimation under non-identifiability that we can still learn from these examples.

Looking at the results, in Appendix 1, we can note that in both experiments the 95% intervals always cover the true value, and that in each case they remain of roughly the same length despite a rather dramatic increase in the population size (thus sample size). This (admittedly wasteful) success at containing the true value was summarily dismissed by ASW as a consequence of the "wide tails" of the posterior distribution, and presented as an indication of the procedure becoming unreliable due to non-identifiability.

My reading is different. First, we should note that the fact that this dataset was generated from a (trivially) non-identifiable model means that ***the data does not contain the necessary information*** for consistent estimation of the population size. This is an unavoidable limitation. This, however, does not necessarily mean that any value will be equally likely *a posteriori*; it just means that there might exist a set of potential values whose plausibility cannot be differentiated using data—this set has the potential of sometimes comprise the whole parameter space, but this neither a necessity nor what we observe here. In our results we can note that: (1) most of the posterior probability mass concentrates around a somewhat reduced region, and (2) that such region *contains the true value*. Put together, these facts suggest a more nuanced reading of what the "wide tails" of the distribution might be telling us: it might not be possible to determine the true value of p_0 out of a set of (indistinguishable based on data) candidates; however the estimation procedure is still able to extract *some* information from the data in the form of a concentration of probability mass around two or more candidate values, which include the true value. In this case, the fact that this region does not shrink with increasing N *suggests* that it is already as small as it can be while still containing all the undecidable values; therefore cannot shrink anymore regardless of the amount of data we might have available.

The above reasoning applies to both experimental examples and should, *in my opinion*, be applicable to many practical situations of non-identifiability in CR. Credible intervals or posterior modal regions which contain the true value of p_0 and do not shrink past a certain population size seem to be a feature of LCM CR estimation under certain types of non-identifiability. We can observe another instance of the same phenomenon in Appendix 2. While, in my opinion, this is the most important takeaway from these experiments, as it exemplifies how posterior estimation can still be useful even as stronger statistical properties fail, we should note that this insight was not available to ASW due to their inadequate computational work.

The two experimental cases, however, have important differences. Comparing them can teach us a fair deal about the effects that different forms of non-identifiability can have in CR estimation using RLCMCRs—and perhaps about CR and LCMs in general.

Let us now move our attention to point estimates—always keeping in mind that, because of non-identifiability, they are likely inconsistent estimators of p_0. We note that these point estimates are surprisingly good in example 3a, while in 4a they are not. We can further note that:

- **3a** ($J = 3, K = 2$): This LCM generates data with only 7 observable capture patterns. The estimation of each latent class takes 4 degrees of freedom. Therefore, these data allows the estimation of LCMs of up to one latent class before becoming trivially non-identifiable. One latent class, however, leaves 3 additional spare degrees of freedom which is close to the needed 4 needed to estimate the extra latent class.

- **4a** ($J = 3, K = 3$): This LCM also allows the estimation of up to one latent class, leaving 3 degrees of freedom to spare. However, different from the previous case, these spare degrees of freedom are a very small fraction of the 12 needed to estimate the two remaining latent classes.

These considerations might help explaining the differences in the quality of point estimation between the two experiments. In the first case (3a), even though there are not enough data to allow purely data-based estimates of the 2nd latent class, there are still available 3 out of the 4 needed degrees of freedom to (1) make it clear that the one-class (independence) model is not a good fit and therefore avoid shrinkage to the 1-class model and, (2) partially reconstruct that class without relying *too much* on priors.

Case 4a, with $K = 3$ latent classes, is different. Here we can apply the same reasoning that we used for the first two latent classes of 3a to the first two of the $K = 3$ latent classes (of 4b). However, this leaves us without any data for fitting the third latent class. Moreover, because of this, the shrinkage is likely favoring a representation closer to 2 effective classes, resulting in bias. This situation is similar to the two-list examples, in Section 4.1, where data was only sufficient for fitting exactly one latent class.

Taking all these results together, I conjecture that fitting RLCMCRs to trivially non-identifiable LCMs might still result in reasonably informative posterior distributions of p_0 as long as the generating LCM family is such that there are enough observable cells to match the number of parameters of an LCM with $K - 1$ latent classes, plus at least a fraction an additional one, i.e. $2^J - 1 > (K - 1)(J + 1) - 1$. Investigating if this conjecture holds (or when and how) might lead to a much better understanding of these procedures.

Models with $J < 2K$ but not trivially non-identifiable (4b, 4c) These are the most interesting cases, as they do violate ASW's sufficient condition for identifiability, but are not trivially non-identifiable. Therefore, different from all previous cases, unexpected or strange behaviors *consistent with non-identifiability* can be reasonably hypothesized to be a consequence of the existence of unknown non-identifiable sets similar, but not equal, to those described in part 2 of Theorem A.2.

- Example **4b** uses $J = 4$ lists and $K = 3$ latent classes. This model is not trivially non-identifiable, but it depletes all the available degrees of freedom. This can cause some problems, since RLCMCR has also to figure out the dimensionality of the problem. However, this should only result in posterior distributions more diffuse than what a model that takes K as an input would produce.
- Example **4c** uses $J = 5$ lists and $K = 3$ latent classes. This produces a contingency table with 31 potentially non-empty cells. Therefore, this model has 13 more degrees of freedom than needed for estimating all the 3 latent classes.

These two examples behave similarly: point estimates (posterior medians) have a small bias that decreases with sample size until becoming negligible. The same occurs with credible intervals: their empirical coverage rate is below the nominal 95% (70% and 51% for 4b and 4c respectively) when $N = 2,000$, but improves as sample size increases until becoming close to nominal (92% and 97%) for $N = 100,000$. Credible intervals for 4c shrink as N increases, at a steady pace. In the case of 4b, there is shrinking, but only when going from $N = 2,000$ to $N = 10,000$.

This behavior is intriguing, but does not seem to me to be particularly consistent with identifiability problems. Specifically, the fact that everything seems to improve as sample size increases is not consistent with what I would expect from non-identifiability, which would be to get posterior distributions unable to get rid of regions that do not contain the true value, as the sample size increases (like we observed in the case of trivial non-identifiability).

Instead, what we observe looks to me more like a procedure which is not so efficient under these circumstances, and has a prior-induced bias. Without going into a lengthier investigation, my best *hypothesis* is that low sample sizes do not provide enough information to overcome the strength of the prior-induced shrinkage, resulting in behavior closer to a lower dimensional model. This behavior diminishes as sample size increases and data is able of take over. This hypothesis could be tested by fitting LCMs with a fixed number of latent classes that do not express any *a priori* preference for any of them; e.g. with prior $\boldsymbol{\pi} \sim \text{Dirchlet}(\mathbf{1}_K)$. Another possibility is that the latent-class level prior distributions might *a priori* be favoring small population sizes. This is a phenomenon that I am currently investigating.

4.3 Takeaways

My main takeaway from this exercise is that ASW's claims are not supported by the evidence they have presented. This does not mean that they are necessarily wrong, though. However, not only I believe that their argumentation is rather weak, but also that the actual evidence from their own experiments better supports different conclusions.

5 Appendix 1: Recomputation of Interesting Latent Class Model Simulations

5.1 Example 3a (trivially NI)

Data generation

Data was simulated from a Latent Class model with $K = 2$ latent classes and $J = 3$ lists with the following parameters:

Table 1: Example 3a: Data generation parameters

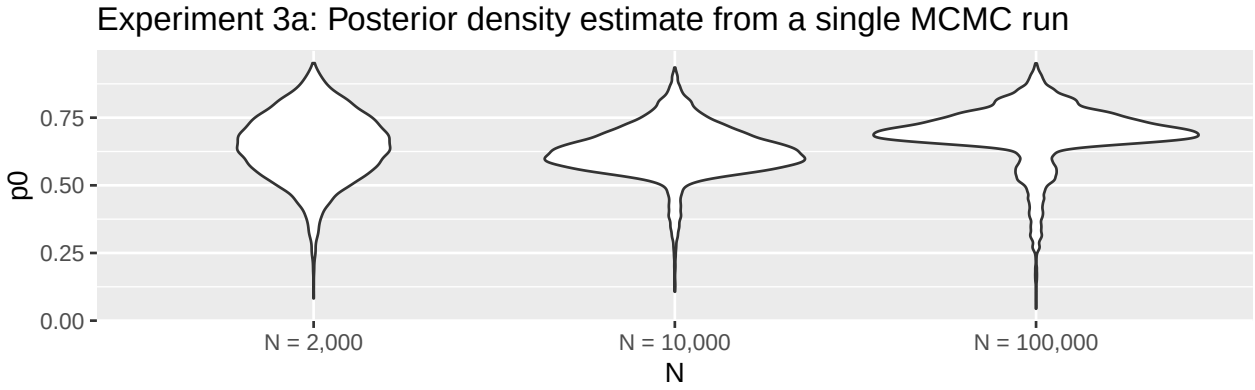
pi	Pr(In List 1)	Pr(In List 2)	Pr(In List 3)
0.9	0.033	0.099	0.132
0.1	0.825	0.759	0.990

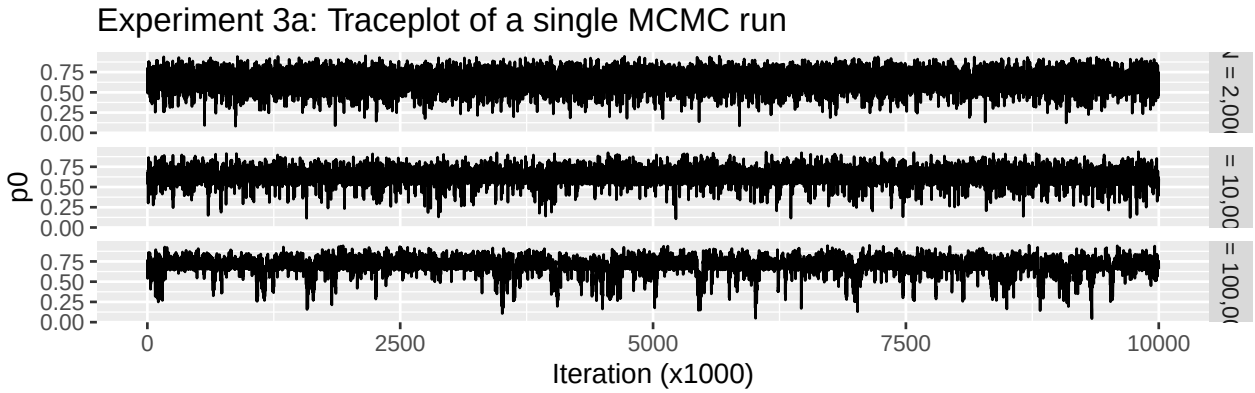
In this case

- The probability of no-capture is $p_0 =$ **0.681**
- $2K > J$, so the model is deemed **non-identifiable by ASW**. However, the total number of parameters (including N) is $K(J + 1) = 8$ and the total number of observable capture patterns is $2^J - 1 = 7$, so *it is also* ***trivially non-identifiable****.*
- $J = 3$ lists allow a maximum of $K = 1$ latent classes before the model becomes trivially non-identifiable, but leave 3 degrees of freedom left. Each additional latent class takes an extra 4 degrees of freedom.

Results:

- **Single run test:**





- **Repeated sampling results**

Table 2: Example 3a: Repeated sampling results

N	n	Mean posterior estimate	95%-CI Coverage	95%-CI length
2,000	638	0.607	100.0%	0.465
10,000	3,197	0.656	100.0%	0.454
100,000	31,941	0.672	100.0%	0.446

5.2 Example 3b (Identifiable)

Data generation

Data was simulated from a Latent Class model with $K = 2$ latent classes and $J = 4$ lists with the following parameters:

Table 3: Example 3b: Data generation parameters

pi	Pr(In List 1)	Pr(In List 2)	Pr(In List 3)	Pr(In List 4)
0.9	0.033	0.099	0.132	0.033
0.1	0.825	0.759	0.990	0.693

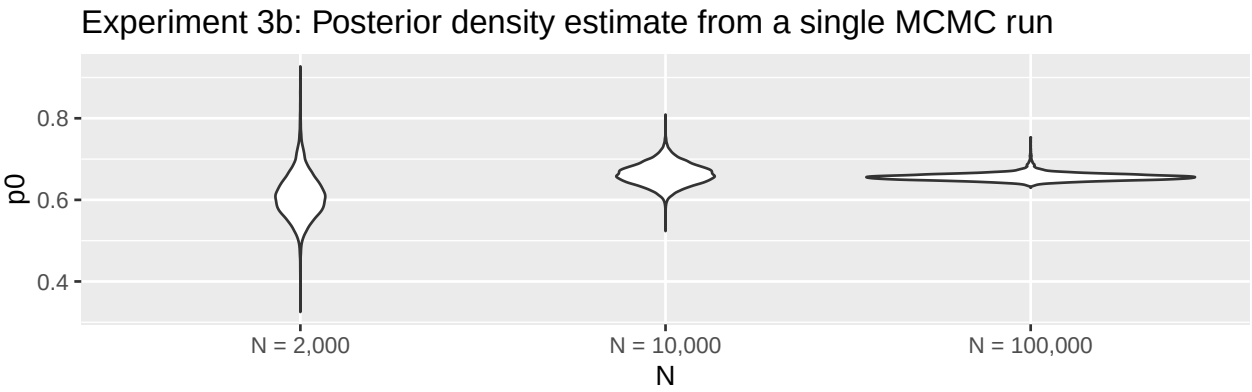
In this case

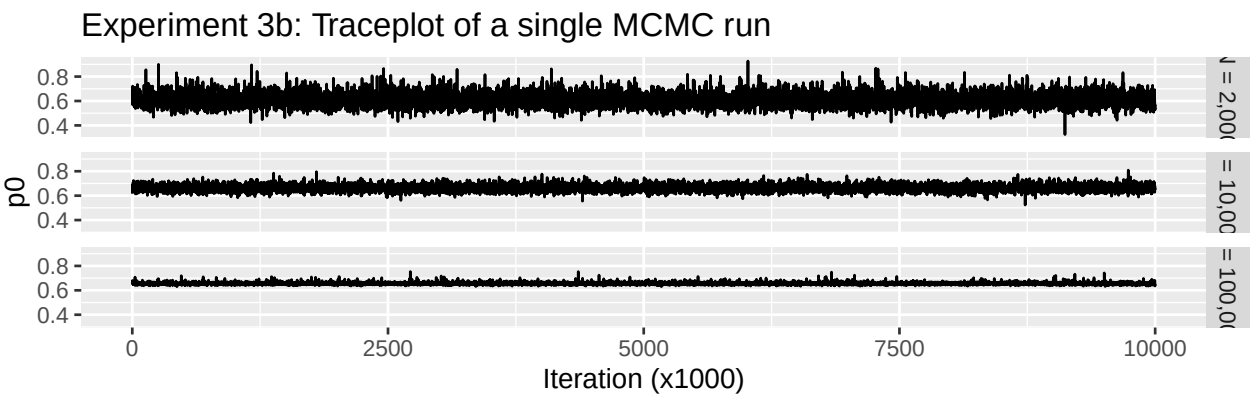
- The probability of no-capture is $p_0 = \mathbf{0.658}$

- $2K \geq J$, therefore this model **identifiable**.
- $J = 4$ lists allow a maximum of $K = 3$ latent classes before the model becomes trivially non-identifiable, and depletes all available degrees of freedom. Each additional latent class takes an extra 5 degrees of freedom.

Results:

- **Single run test:**





- **Repeated sampling results**

Table 4: Example 3b: Repeated sampling results

N	n	Mean posterior estimate	95%-CI Coverage	95%-CI length
2,000	683	0.641	96.5%	0.2010
10,000	3,424	0.660	96.0%	0.0989
100,000	34,195	0.659	97.5%	0.0329

5.3 Example 4a (trivially NI)

Data generation

Data was simulated from a Latent Class model with $K = 3$ latent classes and $J = 3$ lists with the following parameters:

Table 5: Example 4a: Data generation parameters

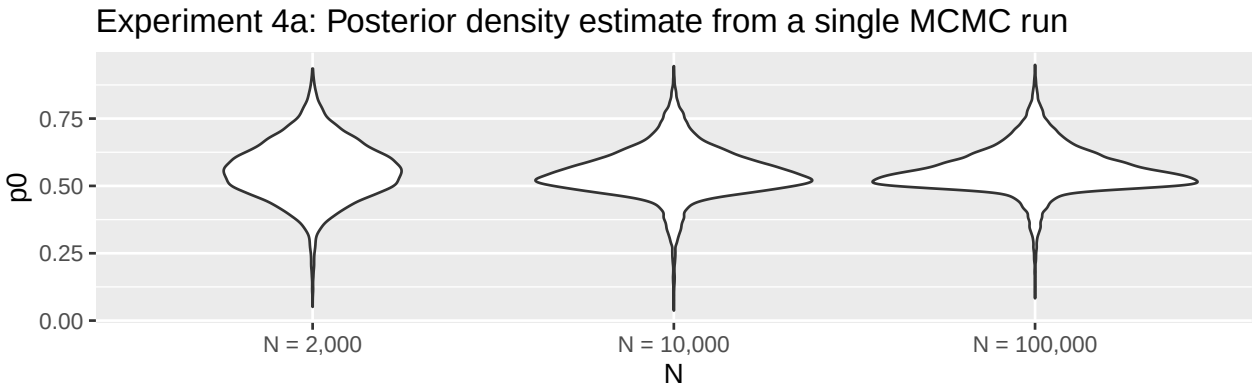
pi	Pr(In List 1)	Pr(In List 2)	Pr(In List 3)
0.7	0.033	0.099	0.132
0.2	0.250	0.200	0.300
0.1	0.825	0.759	0.990

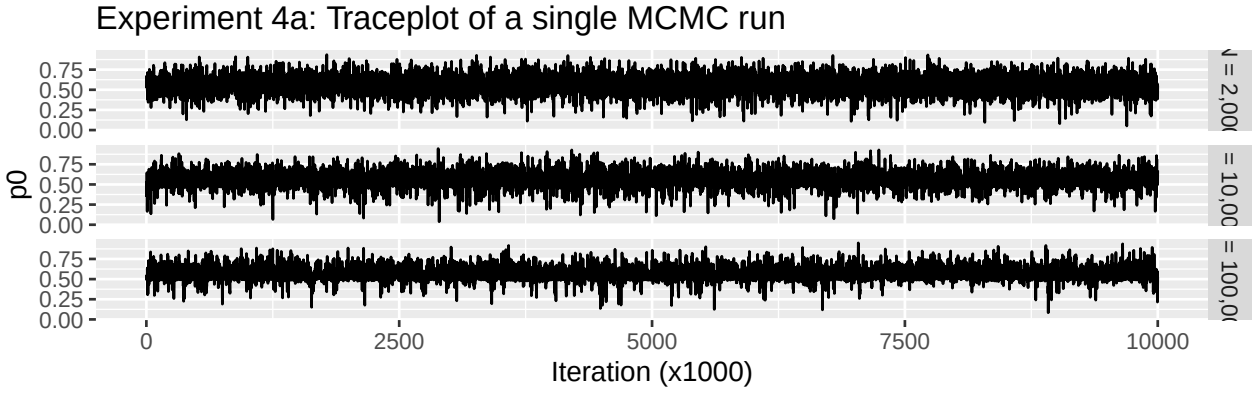
In this case

- The probability of no-capture is $p_0 = \mathbf{0.613}$
- $2K > J$, so the model is deemed **non-identifiable by ASW**. However, the total number of parameters (including N) is $K(J+1) = 12$ and the total number of observable capture patterns is $2^J - 1 = 7$, so *it is also* ***trivially non-identifiable***.
- $J = 3$ lists allow a maximum of $K = 1$ latent classes before the model becomes trivially non-identifiable, but leave 3 degrees of freedom left. Each additional latent class takes an extra 4 degrees of freedom.

Results:

- **Single run test:**





- **Repeated sampling results**

Table 6: Example 4a: Repeated sampling results

N	n	Mean posterior estimate	95%-CI Coverage	95%-CI length
2,000	773	0.529	100.0%	0.433
10,000	3,868	0.550	100.0%	0.409
100,000	38,663	0.556	100.0%	0.399

5.4 Example 4b (J < 2K, but not trivially NI)

Data generation

Data was simulated from a Latent Class model with $K = 3$ latent classes and $J = 4$ lists with the following parameters:

Table 7: Example 4b: Data generation parameters

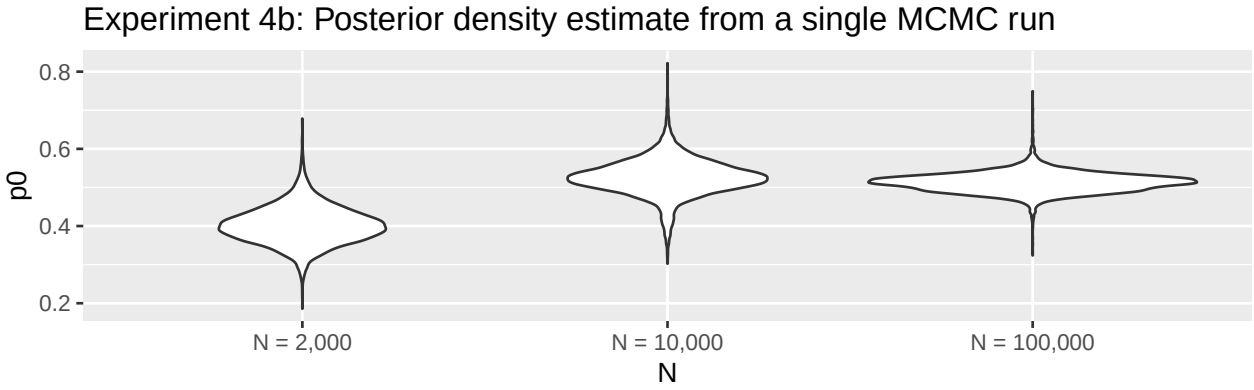
pi	Pr(In List 1)	Pr(In List 2)	Pr(In List 3)	Pr(In List 4)
0.7	0.033	0.099	0.132	0.033
0.2	0.250	0.200	0.300	0.325
0.1	0.825	0.759	0.990	0.693

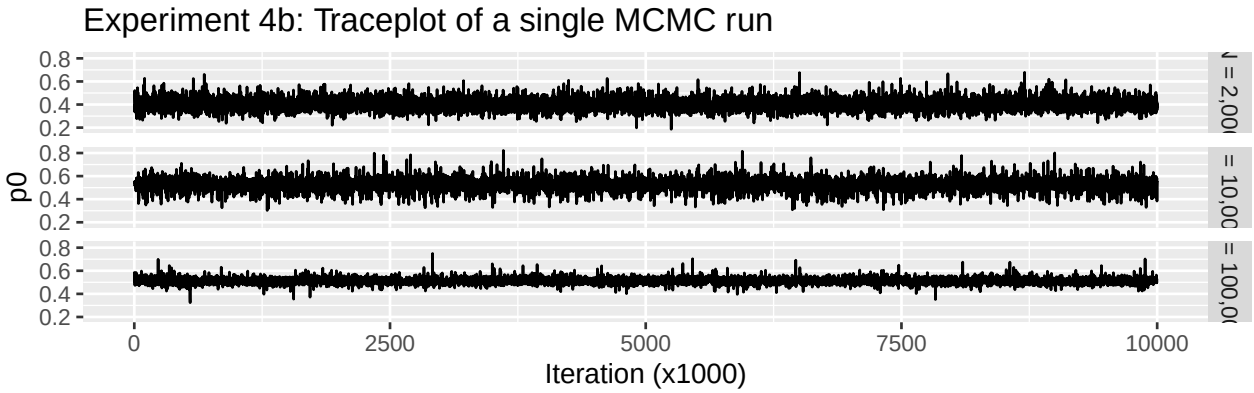
In this case

- The probability of no-capture is $p_0 = \mathbf{0.569}$
- $2K > J$, so the model is deemed **non-identifiable by ASW**, but **not trivially** so, since the total number of parameters (including N) is $K(J + 1) = 15$ and the total number of observable capture patterns is $2^J - 1 = 15$.
- $J = 4$ lists allow a maximum of $K = 3$ latent classes before the model becomes trivially non-identifiable, and depletes all available degrees of freedom. Each additional latent class takes an extra 5 degrees of freedom.

Results:

- **Single run test:**





- **Repeated sampling results**

Table 8: Example 4b: Repeated sampling results

N	n	Mean posterior estimate	95%-CI Coverage	95%-CI length
2,000	861	0.480	70.0%	0.212
10,000	4,316	0.509	68.0%	0.145
100,000	43,145	0.522	92.0%	0.145

5.5 Example 4c (J < 2K, but not trivially NI)

Data generation

Data was simulated from a Latent Class model with $K = 3$ latent classes and $J = 5$ lists with the following parameters:

Table 9: Example 4c: Data generation parameters

pi	Pr(In List 1)	Pr(In List 2)	Pr(In List 3)	Pr(In List 4)	Pr(In List 5)
0.7	0.033	0.033	0.099	0.132	0.033
0.2	0.275	0.250	0.200	0.300	0.325
0.1	0.660	0.825	0.759	0.990	0.693

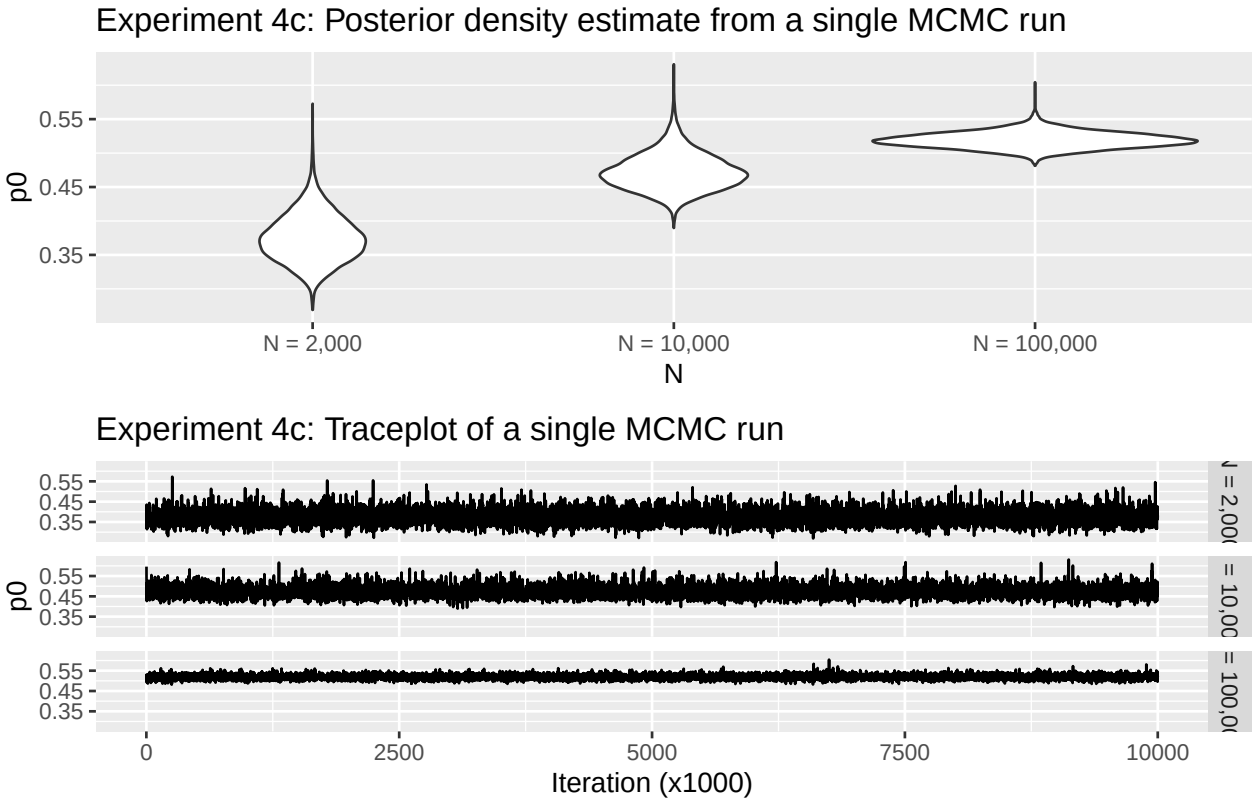
In this case

- The probability of no-capture is $p_0 = \mathbf{0.536}$

- $2K > J$, so the model is deemed **non-identifiable by ASW**, but **not trivially** so, since the total number of parameters (including N) is $K(J+1) = 18$ and the total number of observable capture patterns is $2^J - 1 = 31$.
- $J = 5$ lists allow a maximum of $K = 5$ latent classes before the model becomes trivially non-identifiable, but leave 1 degrees of freedom left. Each additional latent class takes an extra 6 degrees of freedom.

Results:

- **Single run test:**



- **Repeated sampling results**

Table 10: Example 4c: Repeated sampling results

N	n	Mean posterior estimate	95%-CI Coverage	95%-CI length
2,000	926	0.442	51.0%	0.1723
10,000	4,641	0.490	72.5%	0.1304
100,000	46,393	0.531	97.5%	0.0555

6 Appendix 2: An example of posterior estimation under non-identifiability

Here I detail the computational example I discuss in the main text. My objective is to demonstrate how posterior estimation can still provide useful (and in this case, the only available) information about problems where data does not carry enough information for guaranteeing consistent estimation or other desirable repeated-sampling or large sample properties.

I generate data from a LCM with $J = 3$ lists and $K = 2$ latent classes. This example is *trivially non-identifiable* on its own but, like ASW's experiment 3a, it only lacks one degree of freedom. To make it more interesting I have selected parameters according to the second part of Theorem A.2, which guarantees the existence of a different LCM with the same *observable data distribution*, but with a different probability of no-capture. The resulting parameters are:

Table 11: Non-identifiable example: actual parameters

Pr(In List 1)	Pr(In List 2)	Pr(In List 3)	pi
0.4	0.4	0.4	0.8571429
0.8	0.8	0.8	0.1428571

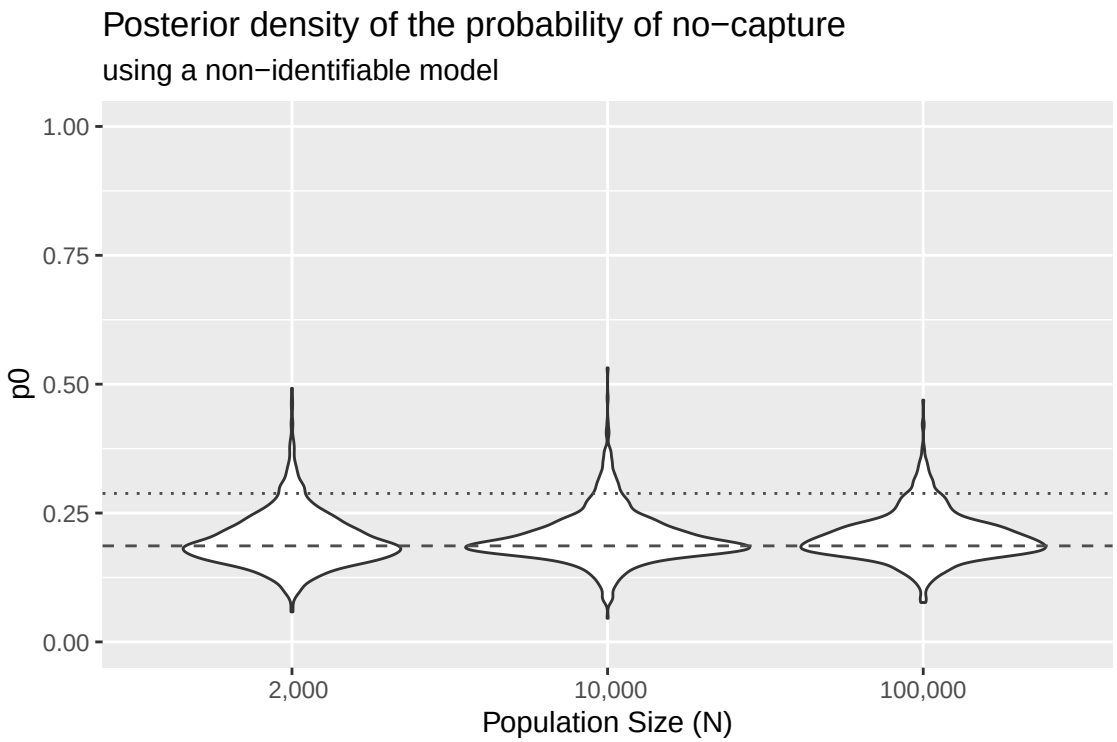
which imply a probability of no-capture of $p_0 = 0.1863$. This model implies the same observable data distribution as

Table 12: Non-identifiable example: wrong parameters

Pr(In List 1)	Pr(In List 2)	Pr(In List 3)	pi
0.2	0.2	0.2	0.5
0.6	0.6	0.6	0.5

which nonetheless implies the probability of no capture $p_0 = 0.288$.

I have simulated CR data from the first model for sample sizes $N = 2,000, 10^4$ and 10^5, and calculated the posterior distribution of p_0.



For completeness I have also performed a repeated-sampling experiment with this specification, similar to ASW's examples. Results are exactly what one would expect from trivial non-identifiability: posterior distributions whose modal regions do not shrink as the sample size increases, but that contain the true value and other undecidable-from-data values.

- **Repeated sampling results**

Table 13: Non-identified model: Repeated sampling results

N	n	Mean posterior estimate	95%-CI Coverage	95%-CI length
2,000	1,627	0.197	100.0%	0.208
10,000	8,141	0.202	100.0%	0.204
100,000	81,370	0.202	100.0%	0.201

References

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Geyer, Charles J. 2011. "Introduction to Markov Chain Monte Carlo." In *Handbook of Markov Chain Monte Carlo*, edited by Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, 45. Boca Raton: CRC Press.

Ishwaran, H., and L. F. James. 2001. "Gibbs Sampling for Stick-Breaking Priors." *Journal of the American Statistical Association* 96: 161–73.

Lazarsfeld, P. F., and N. W. Henry. 1968. *Latent Structure Analysis*. Boston: Houghton Mifflin Co.

Manrique-Vallier, Daniel. 2016. "Bayesian Population Size Estimation Using Dirichlet Process Mixtures." *Biometrics* 72 (4): 1246–54. https://doi.org/10.1111/biom.12502.

Sethuraman, J. 1994. "A Constructive Definition of Dirichlet Priors." *Statistica Sinica* 4: 639–50.