The Central Role of the Identifying Assumption in Population Size Estimation

Serge Aleshin-Guendel[1], Mauricio Sadinle[1], and Jon Wakefield[1,2]

[1]Department of Biostatistics, University of Washington, Seattle, Washington, U.S.A.

[2]Department of Statistics, University of Washington, Seattle, Washington, U.S.A.

Abstract

The problem of estimating the size of a population based on a subset of individuals observed across multiple data sources is often referred to as capture-recapture or multiple-systems estimation. This is fundamentally a missing data problem, where the number of unobserved individuals represents the missing data. As with any missing data problem, multiple-systems estimation requires users to make an untestable identifying assumption in order to estimate the population size from the observed data. If an appropriate identifying assumption cannot be found for a data set, no estimate of the population size should be produced based on that data set, as models with different identifying assumptions can produce arbitrarily different population size estimates—even with identical observed data fits. Approaches to multiple-systems estimation often do not explicitly specify identifying assumptions. This makes it difficult to decouple the specification of the model for the observed data from the identifying assumption and to provide justification for the identifying assumption. We present a re-framing of the multiple-systems estimation problem that leads to an approach which decouples the specification of the observed-data model from the identifying assumption, and discuss how common models fit into this framing. This approach takes advantage of existing software and facilitates various sensitivity analyses. We demonstrate our approach in a case study estimating the number of civilian casualties in the Kosovo war. Code used to produce this manuscript is available at `github.com/aleshing/central-role-of-identifying-assumptions`.

Keywords: Capture-recapture; Missing data; Multiple-systems estimation; Sensitivity analysis.

1 Introduction

Estimating the size of a closed population is a common problem in many fields, including ecology (Otis et al., 1978), epidemiology (Hook and Regal, 1995), official statistics (Anderson and Fienberg, 1999), and human rights (Ball et al., 2002). The available data typically take the form of multiple lists which record information on a subset of individuals in a population. When there exists a mechanism to identify which individuals are the same across lists, multiple-systems estimation (MSE), also known as capture-recapture, provides an approach to estimating the population size based on the overlap of the lists (Bird and King, 2018).

MSE is at its heart a missing data problem, as we do not observe all individuals in the population of interest (see e.g. Fienberg and Manrique-Vallier, 2009; Manrique-Vallier, 2016). As in any missing data problem, MSE requires users to make an untestable identifying assumption about how the observed individuals relate to the unobserved individuals in order to estimate the population size from the observed data. In practice, this means that models with different identifying assumptions can produce arbitrarily different population size estimates, even when the models have identical fits to the observed data. Thus, any identifying assumptions used in an analysis need to be appropriately justified based on the context of the data. If an appropriate identifying assumption can not be found for a data set, no estimate of the population size should be produced based on that data set.

We believe that the central role of specifying the identifying assumption is not sufficiently appreciated, as it is usually conflated with model specification, which involves both making an identifying assumption *and* specifying a model for the observed data. See for example Fienberg (1972) who wrote "... we are assuming that the model which describes the observed data also describes the count of the unobserved individuals. We have no way of checking this assumption," and Manrique-Vallier et al. (2013) who wrote "The arguably most basic assumption in MSE is that the noninclusion of the fully unobserved individuals ... can be represented by the same model that represents the inclusion (and noninclusion) of those we can observe in at least one list. This is a strong and untestable condition."

This conflation of identifying assumption specification and model specification has led practitioners to perform model evaluation by comparing a suite of model fits that are the results of both fundamentally different identifying assumptions and different model specifications for the observed data (see e.g. Sadinle, 2018; Manrique-Vallier et al., 2019; Silverman, 2020). This makes it essentially impossible to disentangle whether differences in inferences are due to differences in identifying assumptions, model specifications for the observed data, or some combination. More importantly, it is rare in these instances for practitioners to provide justification for any of the identifying assumptions being used.

In this article, we propose an approach for MSE that places the identifying assumption front and center in the MSE workflow. We first revisit the framing of MSE as a missing data problem and describe our approach in Section 2. Section 3 reviews two common MSE models—log-linear and latent class models—through our missing data framing. In Section 4 we focus on the identifying assumption associated with log-linear models, and describe how it can be used as a building block for alternative identifying assumptions and sensitivity analyses that examine the impact of the identifying assumption. Finally, in Section 5 we illustrate our approach in a case study of estimating the number of civilian casualties in the Kosovo war.

2 Multiple-Systems Estimation as a Missing Data Problem

2.1 The Data

Suppose we have a closed population of N individuals, of which $n < N$ are observed by one or more of K lists. Let $H = \{0,1\}^K$ denote the possible patterns of inclusion of the individuals in the lists, $H^* = H \setminus \{0\}^K$ denote the possible subsets of lists in which each of the n observed individuals could have been observed, and let $\boldsymbol{x}_i \in H$ denote the subset of lists in which individual i was included. For example, with $K = 3$, $\boldsymbol{x}_i = (0,1,1)$ indicates that individual i was observed in lists 2 and 3, but not list 1.

These data for the N individuals can be gathered into a 2^K contingency table of list overlap, where the cells of the table are indexed by $\boldsymbol{h} \in H$, with counts $n_{\boldsymbol{h}} = \sum_{i=1}^{N} I(\boldsymbol{x}_i = \boldsymbol{h})$. We do not observe the count for cell $\{0\}^K$,

$n_0 := n_{(0,\cdots,0)} = N - n$, which records the number of individuals missing from all lists, so the observed contingency table is incomplete. Let $\boldsymbol{n} = \{n_{\boldsymbol{h}}\}_{\boldsymbol{h} \in H^*}$ denote the counts of the incomplete contingency table. The unobserved cell count n_0, or equivalently the population size N, is the target of inference.

2.2 The Complete-Data Distribution

Under independent and identically distributed (i.i.d.) sampling of individuals by the lists, the 2^K contingency table of counts is multinomially distributed, i.e.

$$\boldsymbol{n}, n_0 \mid N, \boldsymbol{\pi} \sim \textsc{Multinomial}(N, \boldsymbol{\pi}), \tag{1}$$

where $\boldsymbol{\pi} = \{\pi_{\boldsymbol{h}}\}_{\boldsymbol{h} \in H} \in \mathbb{S}^{2^K - 1}$ is a set of cell probabilities, and $\mathbb{S}^d = \{(a_1, \cdots, a_{d+1}) \in \mathbb{R}^{d+1} \mid \sum_{i=1}^{d+1} a_i = 1, a_i > 0 \ \forall i\}$ denotes the d-dimensional probability simplex. We note that this multinomial model, introduced as early as **?**, is a possible simplification of reality, as it does not allow for correlation of individuals' inclusion patterns. We will refer to the model in (1) as the *complete-data distribution*, for which the evaluation relies on knowing the complete 2^K contingency table of counts. In general, the parameter space for this model will be some subset of $\Theta = \{N, \boldsymbol{\pi} \mid N \in \mathbb{N}, \boldsymbol{\pi} \in \mathbb{S}^{2^K - 1}\}$, which we will refer to as the *complete-data parameterization*. As shown in Web Appendix A, when individuals are not i.i.d. sampled, but are sampled independently with cell probabilities drawn i.i.d. from some mixing distribution on $\mathbb{S}^{2^K - 1}$, we also arrive at the model in (1). This is the case for common models for heterogeneity such as the M_h and M_{th} models of Otis et al. (1978). Because common models for heterogeneity reduce to (1), in the rest of this article we will view the cell probabilities as being marginal of any possible heterogeneity mechanisms.

2.3 Decomposing the Complete-Data Distribution

It is instructive to decompose the complete-data distribution as

$$p(\boldsymbol{n}, n_0 \mid N, \boldsymbol{\pi}) = N! \prod_{\boldsymbol{h} \in H} \frac{\pi_{\boldsymbol{h}}^{n_{\boldsymbol{h}}}}{n_{\boldsymbol{h}}!} = L_1(N, \pi_0 \mid n) L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}), \tag{2}$$

with $L_1(N, \pi_0 \mid n) = \binom{N}{n} \pi_0^{N-n} (1 - \pi_0)^n$ and $L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) = n! \prod_{\boldsymbol{h} \in H^*} \tilde{\pi}_{\boldsymbol{h}}^{n_{\boldsymbol{h}}} / n_{\boldsymbol{h}}!$, where $\pi_0 := \pi_{(0,\cdots,0)} = 1 - \sum_{\boldsymbol{h} \in H^*} \pi_{\boldsymbol{h}}$ is the probability of being missing from every list, and $\tilde{\pi}_{\boldsymbol{h}} = \frac{\pi_{\boldsymbol{h}}}{1 - \pi_0} = \frac{\pi_{\boldsymbol{h}}}{\sum_{\boldsymbol{h}' \in H^*} \pi_{\boldsymbol{h}'}}$ is the probability of being observed

in the subset of the lists $\boldsymbol{h}$ conditional on being observed in at least one list. L_1 is a binomial likelihood for n, which has been well studied in the related binomial N problem literature (see e.g. Rukhin, 1975). L_2 is a multinomial likelihood for the observed data $\boldsymbol{n}$ conditional on their sum n, referred to as the *conditional likelihood* (Fienberg, 1972). We will refer to π_0 as the *unobserved cell probability* and to $\tilde{\boldsymbol{\pi}}$ as the *observed cell probabilities*. This decomposition hints at an alternative to the complete-data parameterization Θ, $\Theta^* = \{N, \pi_0, \tilde{\boldsymbol{\pi}} \mid N \in \mathbb{N}, \pi_0 \in (0,1), \tilde{\boldsymbol{\pi}} \in \mathbb{S}^{2^K-2}\}$, which we will refer to as the *observed-data parameterization*. The two parameterizations are equivalent, so we will work with whichever is more convenient for exposition.

2.4 Identifiability

Before performing inference in a statistical model, it is important to check that the model is identifiable. For $\theta \in \Theta^*$, let P_θ denote the complete-data distribution at the set of parameters θ. Consider the following standard definition of identifiability:

Definition 1. *The statistical model* $\mathcal{P}_\Omega = \{P_\theta \mid \theta \in \Omega \subset \Theta^*\}$ *is* ***identifiable*** *if* $\forall \theta_1, \theta_2 \in \Omega$, $P_{\theta_1} = P_{\theta_2}$ *implies that* $\theta_1 = \theta_2$. *Equivalently,* $\mathcal{P}_\Omega$ *is identifiable if* $\forall \theta_1 = \{N, \pi_0, \tilde{\boldsymbol{\pi}}\}, \theta_2 = \{N', \pi_0', \tilde{\boldsymbol{\pi}}'\} \in \Omega$, $L_1(N, \pi_0 \mid n) L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) = L_1(N', \pi_0' \mid n) L_2(\tilde{\boldsymbol{\pi}}' \mid \boldsymbol{n})$ $\forall \boldsymbol{n}$ *implies that* $\theta_1 = \theta_2$.

One can show that the unrestricted model $\mathcal{P}_{\Theta^*}$ is identifiable. Since the goal is to estimate N, sufficiency might lead one to try to estimate N and π_0 in the unrestricted model based solely on the binomial likelihood for n. Examining the likelihood surface for a given n, one finds a maximum at $N = n$ and $\pi_0 = 0$, with a ridge centered along the set $\{N \in \mathbb{N}, \pi_0 \in (0,1) \mid N(1-\pi_0) \approx n\}$ that monotonically decreases as N increases. In Figure 1 we plot this surface when $n = 100$. There is a fundamental problem in that two parameters are being estimated with one data point, which makes it impossible to construct an unbiased or consistent estimator of either N or π_0 (DasGupta and Rubin, 2005; Farcomeni and Tardella, 2012). Thus the standard definition of identifiability is misleading in this setting, as it does not necessarily imply that the parameters are estimable in any traditional sense.

We will instead use the following alternative definition of identifiability specific to MSE (Link, 2003; Holzmann

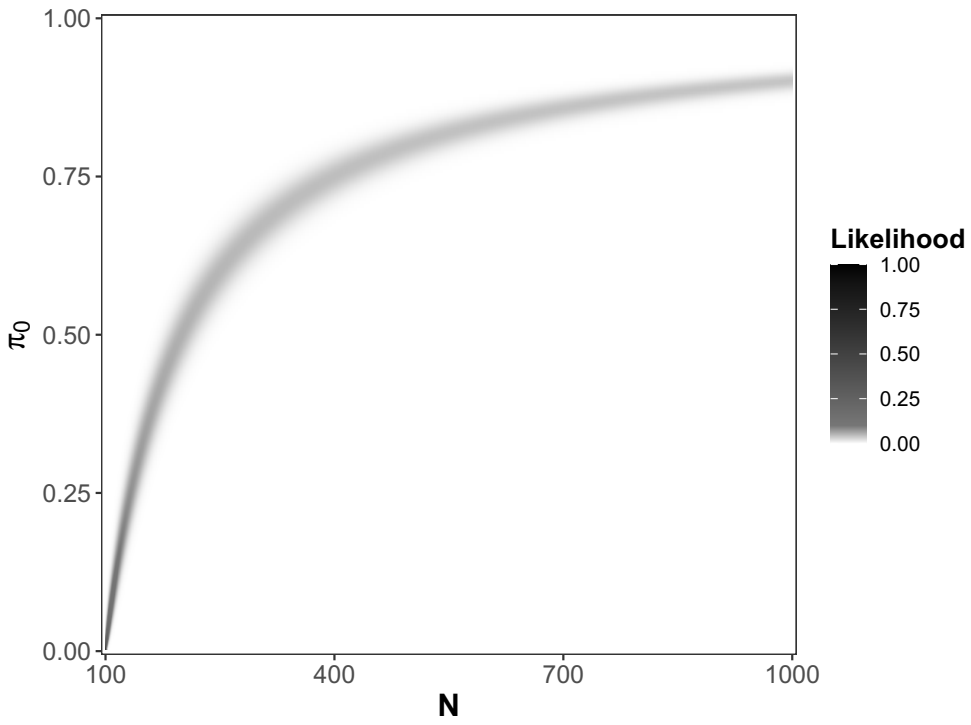

Figure 1: Likelihood surface of L_1 when $n = 100$.

et al., 2006):

Definition 2. *The statistical model* $\mathcal{P}_\Omega$ *is* ***conditionally identifiable*** *if* $\forall \theta_1 = \{N, \pi_0, \tilde{\boldsymbol{\pi}}\}, \theta_2 = \{N', \pi_0', \tilde{\boldsymbol{\pi}}'\} \in \Omega$, $L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) = L_2(\tilde{\boldsymbol{\pi}}' \mid \boldsymbol{n})\ \forall \boldsymbol{n}$ *implies that* $\pi_0 = \pi_0'$.

In a conditionally identifiable model, the conditional likelihood, L_2, identifies the unobserved cell probability, π_0. Clearly the unrestricted model $\mathcal{P}_{\Theta^*}$ is not conditionally identifiable. Standard identifiability of the multinomial conditional likelihood tells us that we can equivalently state Definition 2 as follows: the statistical model $\mathcal{P}_\Omega$ is conditionally identifiable if $\forall \theta_1 = \{N, \pi_0, \tilde{\boldsymbol{\pi}}\}, \theta_2 = \{N', \pi_0', \tilde{\boldsymbol{\pi}}'\} \in \Omega$, $\tilde{\boldsymbol{\pi}} = \tilde{\boldsymbol{\pi}}'$ implies that $\pi_0 = \pi_0'$. Thus for a conditionally identifiable model, there exists a function $\mathcal{T} \colon \tilde{T} \to (0, 1)$ that maps observed cell probabilities, $\tilde{\boldsymbol{\pi}}$, to unobserved cell probabilities, π_0, where $\tilde{T} \subset \mathbb{S}^{2^K - 2}$. When the domain $\tilde{T}$ of this function is not equal to $\mathbb{S}^{2^K - 2}$, this restricts the set of possible values for $\tilde{\boldsymbol{\pi}}$ in the model to $\tilde{T}$. Any extra assumptions in the model involving $\tilde{\boldsymbol{\pi}}$ can then further restrict the set of possible values for $\tilde{\boldsymbol{\pi}}$ in the model to a set $\tilde{S} \subset \tilde{T}$. Thus conditionally identifiable

models take the form $\mathcal{P}_\Omega$, where $\Omega = \{N, \pi_0, \tilde{\boldsymbol{\pi}} \mid N \in \mathbb{N}, \pi_0 = \mathcal{T}(\tilde{\boldsymbol{\pi}}), \tilde{\boldsymbol{\pi}} \in \tilde{S}\}$.

When a model is not conditionally identifiable, we have no guarantees for when the parameters are estimable in any traditional sense. In particular, non-identifiability precludes consistent estimation as "there will be uncertainty in parameter estimates that is not washed out as more data are collected" (Linero, 2017). If a model $\mathcal{P}_\Omega$ is conditionally identifiable, all parameters of the model can be consistently estimated (Sanathanan, 1972). However, we emphasize that the data needs to have been generated by a distribution in the model $\mathcal{P}_\Omega$ for the parameters to be consistently estimable. In other words, in order to estimate the population size N, we need to assume a functional relationship, $\mathcal{T}$, between the observed cell probabilities $\tilde{\boldsymbol{\pi}}$ and the unobserved cell probability π_0. This is the main idea behind MSE.

2.5 Missing Data

The framing in the previous section is motivated by our treatment of MSE as a missing data problem. The decomposition in (2) is related to the decomposition in the missing data literature of the complete-data distribution into the *extrapolation distribution* and the *observed-data distribution* (Hogan and Daniels, 2008). The extrapolation distribution captures how to extrapolate to the missing data given the observed data, which in our context corresponds to L_1. The observed-data distribution, as the name indicates, is the distribution of the observed data, which in this context corresponds to L_2. Following the analogy of the missing data literature, by restricting ourselves to models of the form $\mathcal{P}_\Omega$, where $\Omega = \{N, \pi_0, \tilde{\boldsymbol{\pi}} \mid N \in \mathbb{N}, \pi_0 = \mathcal{T}(\tilde{\boldsymbol{\pi}}), \tilde{\boldsymbol{\pi}} \in \tilde{S}\}$, one is making an *identifying assumption*, $\mathcal{T}$, about how $\tilde{\boldsymbol{\pi}}$ relates to π_0 in order to identify π_0.

The observed-data distribution is restricted when the set of possible values for the observed cell probabilities, $\tilde{S}$, is not equal to $\mathbb{S}^{2^K-2}$. Based on standard properties of the multinomial conditional likelihood, restrictions on the observed-data distribution are assumptions that are testable from the data. As noted in the previous section, these restrictions could be due to the domain, $\tilde{T}$, of the identifying assumption (see Section 4.3 for an example), or due to extra modeling assumptions for the observed cell probabilities, $\tilde{\boldsymbol{\pi}}$ (see Section 3.1 for an example). This motivates the following definition (see Chapter 8 of Hogan and Daniels, 2008):

Definition 3. *A model* $\mathcal{P}_\Omega$*, where* $\Omega = \{N, \pi_0, \tilde{\boldsymbol{\pi}} \mid N \in \mathbb{N}, \pi_0 = \mathcal{T}(\tilde{\boldsymbol{\pi}}), \tilde{\boldsymbol{\pi}} \in \tilde{S}\}$*, is nonparametric identified when* $\tilde{S} = \tilde{T} = \mathbb{S}^{2^K-2}$*, i.e. the observed-data distribution is not restricted by the model.*

2.6 Our Approach to Multiple-Systems Estimation

In the MSE literature, previous work has been concerned with determining *when* certain models are conditionally identified (see e.g. Link, 2003; Holzmann et al., 2006). Here we are concerned with determining both when *and how* models are conditionally identified. Since the validity of our inferences rests on the untestable identifying assumption and any restrictions on the observed-data distribution being correct, we would like to know what identifying assumption we are actually making so we can determine whether or not the assumption is plausible in a given context. Thus, in this article our approach to MSE will be to use conditionally identified models that are based on *explicitly specified* identifying assumptions. Additionally, to make as few testable assumptions as possible, we will use models where the observed-data distribution is only possibly restricted by the identifying assumption (i.e. $\tilde{S} = \tilde{T}$).

Given such a conditionally identified model, our approach to MSE is agnostic to the inferential framework used, so one can perform inference for N in a frequentist or Bayesian framework. In Web Appendix B, we outline how computation, including sensitivity analyses probing the identifying assumption as we will describe in Section 4, can be carried out in either framework using existing software.

In the rest of this article, we examine the identifying assumptions (and sometimes lack thereof) associated with commonly used MSE models, and propose a new family of identifying assumptions. While these identifying assumptions may be useful in some applications, *there is no one-size-fits-all solution.* In practice, the use of identifying assumptions should be accompanied by appropriate justification based on the context of the data. However, in some applications none of the identifying assumptions discussed in this article will be appropriate for the data at hand. There is no default identifying assumption that practitioners can fall back on, and so in these scenarios *no estimate of the population size should be produced based on the data at hand.* Such a scenario is clearly unsatisfactory, and thus it is an important task for researchers in the field of MSE to develop new explicit identifying

assumptions, so that practitioners are able to select identifying assumptions appropriate for their applications.

3 Log-Linear and Latent Class Models

In this section we describe two commonly used models, which we use to demonstrate the drawbacks of using models that either place unnecessary restrictions on the observed-data distribution or that are not based on explicit identifying assumptions.

3.1 Log-Linear Models

For $\boldsymbol{h} \in H^*$, let h_k denote the kth element of $\boldsymbol{h}$. Any set of cell probabilities, $\boldsymbol{\pi} \in \mathbb{S}^{2^K-1}$, can be represented as $\pi_{\boldsymbol{h}} = \mu_{\boldsymbol{h}} / \sum_{\boldsymbol{h}' \in H} \mu_{\boldsymbol{h}'}$, where $\log(\mu_{\boldsymbol{h}}) = \sum_{\boldsymbol{h}' \in H^*} \lambda_{\boldsymbol{h}'} \prod_{k=1}^{K} h_k^{h'_k}$, for some set of log-linear parameters $\boldsymbol{\lambda} = \{\lambda_{\boldsymbol{h}}\}_{\boldsymbol{h} \in H^*} \in \mathbb{R}^{2^K-1}$. This leads to the *log-linear parameterization* $\Theta_{LL} = \{N, \boldsymbol{\lambda} \mid N \in \mathbb{N}, \boldsymbol{\lambda} \in \mathbb{R}^{2^K-1}\}$. Note that under this parameterization, there is no $\lambda_{(0,\cdots,0)}$, so that $\mu_{(0,\cdots,0)} = 1$.

For cells in the incomplete table $\boldsymbol{h} \in H^*$ such that $\sum_{k=1}^{K} h_k = 1$ we refer to $\lambda_{\boldsymbol{h}}$ as a main effect; for $\boldsymbol{h} \in H^*$ such that $\sum_{k=1}^{K} h_k = \ell > 1$ we refer to $\lambda_{\boldsymbol{h}}$ as an ℓ-way interaction. The main effects and interactions all have interpretations as log ratios of certain cross-product ratios (see e.g. Chapter 2 of Bishop et al., 1975). Of particular interest is the K-way, or highest-order, interaction $\lambda_{\mathbf{1}}$, where $\mathbf{1} := (1, \cdots, 1)$, for which we have the relationship $\prod_{\boldsymbol{h} \in H} \pi_{\boldsymbol{h}}^{I_{odd}(\boldsymbol{h})} / \prod_{\boldsymbol{h} \in H} \pi_{\boldsymbol{h}}^{I_{even}(\boldsymbol{h})} = \exp\{(-1)^{K+1} \lambda_{\mathbf{1}}\}$, where $I_{odd}(\boldsymbol{h}) = I(\sum_{k=1}^{K} h_k \text{ is odd})$ and $I_{even}(\boldsymbol{h}) = I(\sum_{k=1}^{K} h_k \text{ is even})$, using the convention that 0 is even. This notation differs from Bishop et al. (1975) as we index the complete table by $H = \{0, 1\}^K$ rather than $\{2, 1\}^K$.

The model $\mathcal{P}_{\Theta_{LL}}$ is equivalent to the unrestricted model $\mathcal{P}_{\Theta}$, so we need to restrict Θ_{LL} to identify the unobserved cell probability π_0. It is standard in this scenario to set $\lambda_{\mathbf{1}} = 0$, so that there is no highest-order interaction in the model. Referring to the resulting parameter space as Ω_{LL}, we would like to understand the identifying assumption made by the *saturated model* $\mathcal{P}_{\Omega_{LL}}$. In Web Appendix C, we show $\mathcal{P}_{\Omega_{LL}}$ is nonparametric identified and that this no-highest-order interaction (NHOI) assumption corresponds to the explicit identifying assumption $\mathcal{T}(\tilde{\boldsymbol{\pi}}) = (\tilde{\Pi}_{odd} / \tilde{\Pi}_{even}) / (1 + \tilde{\Pi}_{odd} / \tilde{\Pi}_{even})$, where $\tilde{\Pi}_{odd} = \prod_{\boldsymbol{h} \in H^*} \tilde{\pi}_{\boldsymbol{h}}^{I_{odd}(\boldsymbol{h})}$ and $\tilde{\Pi}_{even} = \prod_{\boldsymbol{h} \in H^*} \tilde{\pi}_{\boldsymbol{h}}^{I_{even}(\boldsymbol{h})}$, which we

discuss in more detail in Section 4.

In practice there is an emphasis on achieving low variance estimates of the log-linear parameters and, consequentially, N. To this end, rather than just setting the highest-order interaction to zero and using the saturated model, it is common to further restrict the model and set other interactions to zero. This is the case, for example, when restricting to decomposable graphical models (Madigan and York, 1997), or when only including main effects and 2-way interactions (Silverman, 2020), which can be hard to justify in practice (see e.g. Dellaportas and Forster, 1999; Whitehead et al., 2019). This restricts the observed-data distribution, so that we are making a testable assumption that, in addition to the untestable identifying assumption, must be correct in order for inferences to be valid. The hope is that by specifying a model with fewer parameters, the resulting estimates will have lower variance if the chosen restricted model generated the data. However, if the chosen restricted model did not generate the data, estimates of N can be arbitrarily biased, and more generally can have arbitrarily poor frequentist properties (Regal and Hook, 1991; Whitehead et al., 2019).

This is a classic bias-variance trade off, which has been acknowledged since the seminal work of Fienberg (1972) (edited to match our notation): "In analyzing multiple recapture census data our aim is to fit the incomplete 2^K table by a log linear model with the fewest possible parameters, since the fewer parameters in an 'appropriate' model for estimating n_0, the smaller the variance of the estimate. Thus it is not a good practice simply to use the saturated model. On the other hand, if we use a model with too few parameters, we introduce a bias into our estimate of population size that can possibly render the variance formulae of the next section meaningless." Unlike Fienberg (1972), we believe there is a clear route to take if one is using the NHOI assumption, in line with our approach described in Section 2.6: make as few testable assumptions as possible (i.e. use the saturated model $\mathcal{P}_{\Omega_{LL}}$) in the hopes of not being arbitrarily biased because of incorrect restrictions on the observed data distribution. If one does wish to produce lower variance estimators, we discuss in Web Appendix B how regularization can be used to reduce the variance of estimates, at the cost of increasing the bias of estimates, and some difficulties associated with using regularized estimators.

3.2 Latent Class models

Latent class models (LCMs) are typically motivated as models of multivariate categorical data that capture individual heterogeneity when the population can be stratified into J classes, where lists sample individuals independently within each class (Haberman, 1979; Manrique-Vallier, 2016). Thus they are so-called M_{th} models as described in Web Appendix A (Otis et al., 1978). Corollary 1 of Dunson and Xing (2009) shows that for any set of cell probabilities $\boldsymbol{\pi} \in \mathbb{S}^{2^K-1}$, there exists some $J < \infty$ such that $\boldsymbol{\pi}$ can be represented as a J-class latent class model, i.e. $\pi_{\boldsymbol{h}} = \sum_{j=1}^{J} \nu_j \prod_{k=1}^{K} q_{jk}^{h_k}(1-q_{jk})^{1-h_k}$, where $\boldsymbol{\nu} = (\nu_1, \cdots, \nu_J)$ are class membership probabilities, and $\boldsymbol{q} = \{q_{jk}\}_{j=1,k=1}^{J,K}$ are class specific observation probabilities for each list. This leads to the *latent class model parameterization* $\Theta_{LCM} = \{N, \boldsymbol{\nu}, \boldsymbol{q}, J \mid N \in \mathbb{N}, \boldsymbol{\nu} \in \mathbb{S}^{J-1}, \boldsymbol{q} \in (0,1)^{J\times K}, J \in \mathbb{N}\}$. As $\mathcal{P}_{\Theta_{LCM}}$ is equivalent to the unrestricted model $\mathcal{P}_{\Theta}$, we need to restrict Θ_{LCM} to identify the unobserved cell probability π_0. It is common to fix the number of latent classes, J, in advance, to arrive at the the restricted parameterization $\Omega_{LCM,J} = \{N, \boldsymbol{\nu}, \boldsymbol{q} \mid N \in \mathbb{N}, \boldsymbol{\nu} \in \mathbb{S}^{J-1}, \boldsymbol{q} \in (0,1)^{J\times K}\}$.

In Web Appendix A we show that $\mathcal{P}_{\Omega_{LCM,J}}$ is conditionally identified if and only if $2J \leq K$. However, when $\mathcal{P}_{\Omega_{LCM,J}}$ *is* conditionally identified we do not know what explicit identifying assumption is being made or whether the model is nonparametric identified. A recent development in MSE is the use of LCMs with J large enough that $2J > K$ (Manrique-Vallier, 2016). Such LCMs with too many latent classes (i.e. $2J > K$) suffer from the opposite problem of log-linear models: rather than making too many assumptions, and hence restricting the observed-data distribution, so few assumptions are being made that the model is not conditionally identified. In Web Appendix D we show through a variety simulations that this is a practically relevant problem, as we have no guarantees for when estimates based on non-identified models are going to be accurate.

4 Revisiting Log-Linear Models and Their Identifying Assumptions

In this section we revisit the NHOI identifying assumption associated with log-linear models and discuss its role in our framing of MSE. We then describe how this assumption can be used as a building block for alternative identifying assumptions.

4.1 The No-Highest-Order Interaction Assumption

The NHOI assumption introduced in Section 3.1 can be interpreted as follows: for any given subset of $K-1$ lists, appearing in all $K-1$ lists is not associated with appearing or not appearing in the Kth list. Here the meaning of "associated with" changes as the number of lists K changes. When $K=2$ we are assuming that the odds of appearing in list 1 conditional on appearing in list 2 is equal to the odds of appearing in list 1 conditional on not appearing in list 2, and thus the lists are independent: $\pi_{(1,0)}/\pi_{(0,0)} = \pi_{(1,1)}/\pi_{(0,1)}$. When $K=3$ we are assuming that the odds ratio for lists 1 and 2 conditional on appearing in list 3 is equal to the odds ratio for lists 1 and 2 conditional on not appearing in list 3: $\pi_{(1,1,1)}\pi_{(0,0,1)}/(\pi_{(1,0,1)}\pi_{(0,1,1)}) = \pi_{(1,1,0)}\pi_{(0,0,0)}/(\pi_{(1,0,0)}\pi_{(0,1,0)})$. When $K=4$ we assume that certain ratios of odds ratios are equal, and so on for larger K.

As discussed in Section 2.6, in order to use the NHOI assumption in a given application, we need to be able to determine whether or not it is plausible. Odds and odds ratios are commonly used in statistics (Bishop et al., 1975), and thus the NHOI assumption may be of use when there are $K=2$ or $K=3$ lists. However, higher order measures of association like ratios of odds ratio are more obscure and hard to interpret, which makes the NHOI assumption difficult to use when there are more than $K=3$ lists. This difficulty compounds when considering sensitivity analyses as we explain in the next section.

4.2 Sensitivity Analyses for the No-Highest-Order Interaction Assumption

Sensitivity analyses aim to gauge how sensitive inferences are to untestable assumptions, and are an important part of missing data workflows (see Chapter 9 of Hogan and Daniels, 2008). The NHOI assumption facilitates sensitivity analyses based on varying the highest-order interaction across a range of non-zero values. In particular, when fixing $\xi = \exp\{(-1)^{K+1}\lambda_{\mathbf{1}}\} \in \mathbb{R}^+$, we show in Web Appendix C that we arrive at the explicit identifying assumption $\mathcal{T}(\tilde{\boldsymbol{\pi}}) = (\tilde{\Pi}_{odd}/\tilde{\Pi}_{even})/(\xi + \tilde{\Pi}_{odd}/\tilde{\Pi}_{even})$. This generalizes the two list sensitivity analyses of Lum and Ball (2015) and Gerritse et al. (2015). Under this identifying assumption, rather than assuming certain measures of association are equal, we are assuming one measure is ξ times another. For example, when $K=2$ we are assuming that the odds of appearing in list 1 conditional on not appearing in list 2 is ξ times the odds of appearing in list 1

conditional on appearing in list 2: $\pi_{(1,0)}/\pi_{(0,0)} = \xi\pi_{(1,1)}/\pi_{(0,1)}$.

In order to perform a meaningful sensitivity analysis, one needs to be able to specify a range of values for the highest-order interaction that are plausible for a given application. Due to our understanding of odds and odds ratios, performing this sort of sensitivity analysis may be possible when there are $K = 2$ or $K = 3$ lists. When considering more than $K = 3$ lists, it can become difficult to even start thinking about whether it is plausible that ξ is less than or greater than 1, let alone determine specific values of ξ that are plausible.

4.3 K'-List Marginal No-Highest-Order Interaction Assumptions

The NHOI assumption can be used as a building block to generate other identifying assumptions. Suppose we can assume that, without loss of generality, the NHOI assumption holds for the first $1 < K' < K$ lists, marginal of the remaining $K - K'$ lists. This leads to a new identifying assumption which in general does not imply that there is no highest-order interaction for all K lists. To introduce this assumption formally we need to introduce some notation. Let $G = \{0,1\}^{K'}$ index the marginal $2^{K'}$ contingency table for the first K' lists and $G^* = G\backslash\{0\}^{K'}$. For a set of cell probabilities, $\boldsymbol{\pi} \in \mathbb{S}^{2^K-1}$, and a given cell in the marginal table, $\boldsymbol{g} \in G$, let $\pi_{\boldsymbol{g}+} = \sum_{\boldsymbol{h}\in H} \pi_{\boldsymbol{h}} I\{(h_1, \cdots, h_{K'}) = \boldsymbol{g}\}$ denote the probability of being observed in cell $\boldsymbol{g}$ of the marginal table implied by $\boldsymbol{\pi}$. Similarly let $\tilde{\pi}_{\boldsymbol{g}+} = \sum_{\boldsymbol{h}\in H^*} \tilde{\pi}_{\boldsymbol{h}} I\{(h_1, \cdots, h_{K'}) = \boldsymbol{g}\}$ and $\tilde{\pi}_{0+} = \sum_{\boldsymbol{h}\in H^*} \tilde{\pi}_{\boldsymbol{h}} I\{(h_1, \cdots, h_{K'}) = (0, \cdots, 0)\}$.

Assuming that the NHOI assumption holds for the first $1 < K' < K$ lists, marginal of the remaining $K - K'$ lists, is equivalent to assuming $\prod_{\boldsymbol{g}\in G} \pi_{\boldsymbol{g}+}^{I_{odd}(\boldsymbol{g})} / \prod_{\boldsymbol{g}\in G} \pi_{\boldsymbol{g}+}^{I_{even}(\boldsymbol{g})} = 1$. In Web Appendix C we show that this K'-list marginal no-highest-order interaction assumption corresponds to the explicit identifying assumption $\mathcal{T}(\tilde{\boldsymbol{\pi}}) = (\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \tilde{\pi}_{0+})/(1 + \tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \tilde{\pi}_{0+})$, where $\tilde{\Pi}_{odd,+} = \prod_{\boldsymbol{g}\in G^*} \tilde{\pi}_{\boldsymbol{g}+}^{I_{odd}(\boldsymbol{g})}$ and $\tilde{\Pi}_{even,+} = \prod_{\boldsymbol{g}\in G^*} \tilde{\pi}_{\boldsymbol{g}+}^{I_{even}(\boldsymbol{g})}$. Further, we can perform sensitivity analyses for this assumption by fixing $\prod_{\boldsymbol{g}\in G} \pi_{\boldsymbol{g}+}^{I_{odd}(\boldsymbol{g})} / \prod_{\boldsymbol{g}\in G} \pi_{\boldsymbol{g}+}^{I_{even}(\boldsymbol{g})} = \xi \in \mathbb{R}^+$. As we show in Web Appendix C, this leads to the explicit identifying assumption

$$\mathcal{T}(\tilde{\boldsymbol{\pi}}) = \frac{\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \xi\tilde{\pi}_{0+}}{\xi + (\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \xi\tilde{\pi}_{0+})}. \tag{3}$$

Models that use the assumption that $\prod_{\boldsymbol{g}\in G} \pi_{\boldsymbol{g}+}^{I_{odd}(\boldsymbol{g})} / \prod_{\boldsymbol{g}\in G} \pi_{\boldsymbol{g}+}^{I_{even}(\boldsymbol{g})} = \xi \in \mathbb{R}^+$ are not nonparametric identified, as the domain of the identifying assumption is $\tilde{T} = \{\tilde{\boldsymbol{\pi}} \in \mathbb{S}^{2^K-2} \mid \tilde{\Pi}_{odd,+}/(\tilde{\Pi}_{even,+}\tilde{\pi}_{0+}) > \xi\}$.

A special case of this identifying assumption was originally suggested in Regal and Hook (1998) as an alternative to the NHOI assumption. They considered a data set consisting of $K = 3$ lists recording cases of spina bifida in upstate New York, where they believed that the assumption that two of the lists were marginally independent (i.e., using the 2-list marginal NHOI assumption) was more plausible than the NHOI assumption. This illustrates that there may be applications where one may be more willing to make marginal assumptions about a subset of K' lists, rather than an assumption involving all K lists. Additionally when there are $K > 3$ lists and $K' = 2$ or $K' = 3$, the K'-list marginal NHOI assumption and its sensitivity analyses are much more straightforward to interpret than the highest-order interaction and its sensitivity analyses, as discussed in Sections 4.1 and 4.2.

For these reasons, we believe that the K'-list marginal NHOI assumption can be useful as an explicit identifying assumption in the toolbox of the MSE practitioner. However, we emphasize here our message from Section 2.6: there are no one-size-fits-all identifying assumptions. Specification of identifying assumptions in practice should be accompanied with appropriate justification based on the context of the data. In Section 5.1 we attempt to provide such a justification for our use of the 2-list marginal NHOI assumption in an application estimating the number of civilian casualties in the Kosovo war.

5 Civilian Casualties in the Kosovo War

In this section we estimate the number of civilian casualties in the Kosovo war between March 20 and June 22, 1999, using data originally analyzed in Ball et al. (2002). The data consist of $K = 4$ lists with $n = 4400$ observed casualties, and are presented in Table 1, reproduced from Section 6 of Ball et al. (2002). Three of the lists were constructed from refugee interviews conducted separately by the American Bar Association Central and East European Law Initiative (ABA), Human Rights Watch (HRW), and the Organization for Security and Cooperation in Europe (OSCE). The fourth list was constructed from exhumation reports conducted on behalf of the International Criminal Tribunal for the Former Yugoslavia (EXH). We refer the reader to Appendix 1 of Ball et al. (2002) for a detailed description of each list.

The Kosovo data was originally analyzed in Ball et al. (2002) under the NHOI assumption, but as we discuss in

Table 1: Kosovo dataset, reproduced from Section 6 of Ball et al. (2002).

	ABA	yes	yes	no	no
	EXH	yes	no	yes	no
HRW	OSCE				
yes	yes	27	32	42	123
yes	no	18	31	106	306
no	yes	181	217	228	936
no	no	177	845	1131	n_0

the next section, we believe the K'-list marginal NHOI assumption is more appropriate. We will analyze the Kosovo data under both assumptions, highlighting the importance of careful specification of the identifying assumption.

5.1 Choice of Identifying Assumption

For our main analysis we will consider two identifying assumptions. The first assumption is the 2-list marginal NHOI assumption described in Section 4.3, where we will assume that the ABA and HRW lists are marginally independent. We believe this assumption is plausible given that "there were no overt efforts by any of the researchers to exclude or include witnesses who had participated in another data collection project" (ABA/AAAS, 2000, p. 40) and that the two lists had similarly extensive geographic reach in their interviews. In particular, ABA conducted interviews in Albania, Macedonia, Kosovo, the United States, and Poland, while HRW conducted interviews in Albania, Macedonia, Kosovo, and Montenegro. ABA only conducted around 10% of its interviews in the United States and Poland, and HRW only conducted 3% of its interviews in Montenegro. Further, within Kosovo, ABA and HRW conducted interviews in similar geographic regions. For more information on where the lists conducted interviews see Appendix 1 of Ball et al. (2002).

The original analysis of the Kosovo data set in Ball et al. (2002) used the NHOI assumption described in Section 3.1. To justify this assumption for the Kosovo data, as we have $K = 4$ lists, we would need to reason about certain ratios of odds ratios being equal, which can be difficult, as discussed in Section 4.1 and further explained

in Web Appendix E. Nevertheless, we will also analyze the Kosovo data using the NHOI assumption to highlight the importance of careful specification of the identifying assumption.

5.2 Inference

For each identifying assumption, our main analysis will present both a frequentist analysis and a Bayesian analysis, using the methods discussed in Web Appendix B, to demonstrate how our proposed approach to MSE is agnostic to the inferential framework used. The Bayesian analysis will use a negative-binomial prior for N and the prior induced for the observed cell probabilities $\tilde{\boldsymbol{\pi}}$ from using the Dirichlet process prior of Manrique-Vallier (2016) for the J class LCM $\Omega_{LCM,J}$, with $J = 10$ and default hyperparameters, as implemented in the `R` package `LCMCR` (see Web Appendix B for further details). In Web Appendix E we perform a prior sensitivity analysis for the Bayesian analyses, exploring the impact of the priors for N and $\tilde{\boldsymbol{\pi}}$ on our estimates of N.

To inform the negative-binomial prior for N, we will rely on two studies that attempted to estimate the number of casualties in the Kosovo war using different data sources than Ball et al. (2002). Spiegel and Salama (2000) estimated there were 12000 casualties with a 95% confidence interval of $[5500, 18300]$. Iacopino et al. (2001) estimated there were 8000 casualties with a 95% confidence interval of $[5800, 10200]$. Using the negative-binomial parameterization given in Table 1 of Web Appendix B, we will use a specification with mean $M = 10000$ (the average of the estimates from the two studies) and overdispersion parameter $a = 1.6$, which places 95% of the prior mass on $[818, 30371]$. This specification is meant to be weakly informative in the sense that the information it incorporates is intentionally weaker than what is available to us, so as to provide a proper alternative to the "noninformative" improper scale prior $p(N) \propto 1/N$ discussed in Web Appendix B (see e.g. Gelman et al., 2017). This prior places mass below the observed sample size of $n = 4400$, as we are not attempting to use the observed data to inform our prior. Practically speaking this does not make a difference, as the prior is effectively truncated to $[n, \infty)$ when performing posterior inference.

5.3 Main Analysis

In Table 2 we present the results from our frequentist and Bayesian analyses under the 2-list marginal NHOI assumption, i.e. assuming marginal independence of the ABA and HRW lists. Assuming marginal independence of the ABA and HRW lists, under a frequentist analysis we estimate there were 9691 civilian casualties, with a 95% confidence interval of $[8074, 11308]$, and under a Bayesian analysis with the chosen priors we estimate there were 9359 civilian casualties, with a 95% credible interval of $[7967, 11059]$. These point estimates and uncertainty intervals from these two analyses are in close agreement. Both of the uncertainty intervals include the point estimate from Iacopino et al. (2001), but not from Spiegel and Salama (2000), and fall within the confidence interval of Spiegel and Salama (2000). Based on the results of the prior sensitivity analysis in Appendix E, the Bayesian analysis is not sensitive to the prior choices for N and $\tilde{\boldsymbol{\pi}}$.

Table 2: Point estimates and 95% uncertainty intervals for N under the 2-list marginal NHOI assumption. For the Bayesian analysis the point estimate is the posterior mean.

	Point Estimate	Uncertainty Interval
Frequentist	9691	[8074, 11308]
Bayesian	9359	[7967, 11059]

In Table 3 we present the results from our frequentist and Bayesian analyses under the NHOI assumption. Under the NHOI assumption, under a frequentist analysis we estimate there were 16941 civilian casualties, with a 95% confidence interval of $[5304, 28579]$, and under a Bayesian analysis with the chosen priors we estimate there were 14071 civilian casualties, with a 95% credible interval of $[9321, 21604]$. The point estimates and uncertainty intervals from these two analyses are in relative agreement. Both of the uncertainty intervals include the point estimate from Spiegel and Salama (2000), and the frequentist confidence interval includes the point estimate. Based on the results of the prior sensitivity analysis in Appendix E, the Bayesian analysis is fairly sensitive to the prior choices for N and $\tilde{\boldsymbol{\pi}}$.

Focusing on point estimates, we see a large difference between the analyses under the two identifying assump-

Table 3: Point estimates and 95% uncertainty intervals for N under the NHOI assumption. For the Bayesian analysis the point estimate is the posterior mean.

	Point Estimate	Uncertainty Interval
Frequentist	16941	[5304, 28579]
Bayesian	14071	[9321, 21604]

tions (besides the uncertainty interval widths being considerably larger under the NHOI assumption). The point estimates under the NHOI assumption are 75% larger for the frequentist analyses (50% larger for the Bayesian analyses) than the point estimates under the 2-list marginal NHOI assumption. If the 2-list marginal NHOI assumption truly holds, as we are inclined to believe based on the justification provided in Section 5.1, an analysis based on using the NHOI assumption produces estimates with a large positive bias for the Kosovo data. This should serve as an illustration of the dangers of using the NHOI assumption (or any other identifying assumption) that can not be justified based on the context of the data. If a practitioner can not find an identifying assumption that is appropriate for their data, no estimate of the population size should be produced based on their data, as there is no one-size-fits-all or default identifying assumption to fall back on. There is a need for researchers to develop new explicit identifying assumptions, so that practitioners do not find themselves in such a scenario.

5.4 A Sensitivity Analysis Probing the 2-List Marginal NHOI Assumption

While we believe that it is plausible that the ABA and HRW lists are marginally independent, we would also like to understand how sensitive our resulting estimates are to realistic violations of the assumption. If this marginal independence was violated, it would likely be the case that the lists are positively dependent and thus population size estimates under marginal independence are downward biased, as is common in human rights applications (see e.g. the discussion in Section 5 of Lum and Ball, 2015). In particular, HRW selected regions in Kosovo to conduct interviews based on reports of human rights violations from refugees and other sources (ABA/AAAS, 2000). Thus it seems possible that a casualty appearing in HRW could be more likely to appear in ABA than a casualty that did not appear in HRW.

We now perform a sensitivity analysis probing the 2-list marginal NHOI assumption. In Web Appendix E, we provide a similar sensitivity analysis probing the NHOI assumption. We will consider models with the identifying assumption (3), varying ξ over $\{0.7, 0.8, 0.9, 1\}$. Thus in each case we are assuming that the odds of appearing in ABA conditional on not appearing in HRW is ξ times the odds of appearing in ABA conditional on appearing in HRW, with $\xi = 1$ corresponding to the 2-list marginal NHOI assumption. For each value of ξ, we will present both a frequentist analysis and a Bayesian analysis, with the Bayesian analysis using the same priors from the main analysis as presented in Section 5.2. In Table 4 we present the results from our frequentist and Bayesian analyses under each identifying assumption.

Table 4: Point estimates and 95% uncertainty intervals for sensitivity analysis probing the 2-list marginal NHOI assumption. For the Bayesian analysis the point estimate is the posterior mean. In this table ξ is a marginal odds ratio, as described in Section 4.3.

	$\xi = 1$	$\xi = 0.9$	$\xi = 0.8$	$\xi = 0.7$
Frequentist	9691 [8074, 11308]	10534 [8738, 12330]	11588 [9568, 13607]	12942 [10636, 15249]
Bayesian	9359 [7967, 11059]	10155 [8607, 12038]	11147 [9419, 13258]	12419 [10451, 14816]

The estimates of the number of casualties N increase as the amount of assumed positive dependence increases, i.e. as ξ decreases, as expected. When $\xi = 0.9$, the point estimates and uncertainty intervals are still largely compatible with the point estimates and uncertainty intervals under marginal independence. Thus our estimates under marginal independence are not sensitive to this small amount of positive dependence. However, this is not still the case under stronger positive dependence. When $\xi = 0.7$, the uncertainty intervals barely overlap with the uncertainty intervals under marginal independence, and further they do not contain the point estimates under marginal independence. While this may seem like cause for concern, we note that these estimates under stronger positive dependence are still within an order of magnitude of the estimates under independence, and all uncertainty intervals in this sensitivity analysis fall within the confidence interval of Spiegel and Salama (2000). We note that the frequentist analysis requires a marginal odds ratio of $\xi \approx 0.51$ to produce a point estimate as large as the point estimate under the NHOI assumption. This is a large amount of positive dependence which casts further doubt on

the plausibility of the NHOI assumption.

6 Discussion

In this article we revisited the framing of MSE as a missing data problem and proposed an approach for MSE that places the identifying assumption front and center in the MSE workflow. As we have emphasized throughout this article, a natural next step is to develop new explicit identifying assumptions, for situations where the identifying assumptions described in Section 4 can not be justified in the context of a given data set. We believe that this is an extremely under-researched problem that will hopefully gain attention with the re-framing of MSE we present in this article.

The presentation of MSE in this article was focused on estimating the size of a single population. When the population can be stratified based on observed covariates, such as location or time, it may be desirable to estimate the population sizes within each strata. In theory, the methodology developed in this article could be applied independently to each strata. However, stratification can lead to sparse contingency tables, which need significant regularization when estimating $\tilde{\boldsymbol{\pi}}$. In this case, it would be desirable to develop observed data models that borrow strength across strata.

A Web Appendix A: Conditional Identifiability in Models for Heterogeneity

The purpose of this appendix is to show how common models for heterogeneity fit into the model described in Section 2.2 of the main text, and to provide results regarding conditional identifiability in a particular family of heterogeneous models. The material presented in Appendices A.2, A.3, A.4, and A.5 previously appeared in the unpublished preprint Aleshin-Guendel (2020).

A.1 Models for Heterogeneity

Consider the following heterogeneous model

$$\begin{aligned} \boldsymbol{\pi}^i &\overset{i.i.d.}{\sim} Q, \\ \boldsymbol{x}_i \mid \boldsymbol{\pi}^i &\overset{ind.}{\sim} \textsc{Categorical}(\boldsymbol{\pi}^i), \end{aligned} \tag{A.1}$$

where $\boldsymbol{\pi}^i = \{\pi^i_{\boldsymbol{h}}\}_{\boldsymbol{h}\in H} \in \mathbb{S}^{2^K-1}$ for $i = 1, \ldots, N$. Under this model each individual has its own set of cell probabilities, $\boldsymbol{\pi}^i$, drawn from some mixing distribution Q on $\mathbb{S}^{2^K-1}$. Working with the heterogeneous model in (A.1) is equivalent, after marginalizing out $\boldsymbol{\pi}^i$, to working with the complete-data distribution in Equation (1) of the main text, where $\boldsymbol{\pi} := \boldsymbol{\pi}_Q = E_Q(\boldsymbol{\pi}^i)$ and E_Q denotes the expectation with respect to the mixing distribution Q. This is a consequence of the data only providing information about the first moment of the mixing distribution. Suppose $\mathcal{Q}$ is a family of mixing distributions on $\mathbb{S}^{2^K-1}$. For $Q \in \mathcal{Q}$, let $\pi_{Q,0}$ denote the induced observed cell probability and $\tilde{\boldsymbol{\pi}}_Q$ denote the induced observed cell probabilities. The parameter space induced by the family $\mathcal{Q}$, as a subset of the observed-data parameterization, can then be written as $\Omega_{\mathcal{Q}} = \{N, \pi_0, \tilde{\boldsymbol{\pi}} \mid N \in \mathbb{N}, \pi_0 = \pi_{Q,0} \text{ and } \tilde{\boldsymbol{\pi}} = \tilde{\boldsymbol{\pi}}_Q \text{ for some } Q \in \mathcal{Q}\}$.

The general heterogeneous model in (A.1) captures common models for heterogeneity, including the M_h and M_{th} models (Otis et al., 1978). The M_{th} model assumes the individual cell probabilities take the form $\pi^i_{\boldsymbol{h}} = \prod_{k=1}^K (q^i_k)^{h_k}(1-q^i_k)^{1-h_k}$, where $(q^i_1, \cdots, q^i_K) \overset{i.i.d.}{\sim} Q$ and Q is a mixing distribution on $(0,1)^K$. Under this model, conditional on an individual's sampling probabilities, $(q^i_1, \cdots, q^i_K)$, each individual is independently sampled by

each list. The M_h model is a submodel of the M_{th} model that assumes that the individual sampling probabilities, $(q_1^i, \cdots, q_K^i)$, are the same for each list, i.e. $q_1^i = \cdots = q_K^i$. Thus the M_h model assumes individuals have the same probability of being sampled by each list. After marginalizing out $\boldsymbol{\pi}^i$, this enforces a symmetry where the probability of appearing in k lists is the same for each subset of k lists. We do not believe this is plausible in human population settings.

A.2 Conditional Identifiability in M_{th} Models

While there exists a literature characterizing identifiability in M_h models (Huggins, 2001; Link, 2003; Holzmann et al., 2006; Link, 2006), no such results exist for M_{th} models. The purpose of this section is to provide a mechanism for verifying whether the M_{th} model $\mathcal{P}_{\Omega_\mathcal{Q}}$ is conditionally identifiable based on moments of the mixing distributions $Q \in \mathcal{Q}$, analogously to the results for M_h models presented in Holzmann et al. (2006).

Before proving the main theorem of this section, we have the following lemma, which tells us that for any mixing distribution Q on $(0,1)^K$, the induced cell probabilities, $\boldsymbol{\pi}_Q$, only depend on Q through its mixed moments.

Lemma A.1. *For any* $\boldsymbol{h} \in H^*$, $\pi_{Q,\boldsymbol{h}} = \sum_{\boldsymbol{h}' \in H^*} c_{\boldsymbol{h},\boldsymbol{h}'} m_{Q,\boldsymbol{h}'}$ *where* $c_{\boldsymbol{h},\boldsymbol{h}'} = (-1)^{\sum_{k=1}^K h'_k - h_k} \prod_{k=1}^K I(h_k \leq h'_k)$ *and* $m_{Q,\boldsymbol{h}'} = E_Q(\prod_{k=1}^K q_k^{h'_k})$.

Proof. For all $\boldsymbol{h} \in H^*$, $\prod_{k=1}^K q_k^{h_k}(1-q_k)^{1-h_k} = \sum_{\boldsymbol{h}' \in H^*} c_{\boldsymbol{h},\boldsymbol{h}'} \prod_{k=1}^K q_k^{h'_k}$ by an application of the multi-binomial theorem (a generalization of the binomial theorem). The result follows from taking the expectation over both sides with respect to Q. □

We can restate Lemma A.1 in matrix form. Letting $\boldsymbol{\pi}_Q^* = (\pi_{Q,\boldsymbol{h}})_{\boldsymbol{h} \in H^*}$ and $\boldsymbol{m}_Q = (m_{Q,\boldsymbol{h}})_{\boldsymbol{h} \in H^*}$, we have that $\boldsymbol{\pi}_Q^* = C\boldsymbol{m}_Q$, where $C = (c_{\boldsymbol{h},\boldsymbol{h}'})_{\boldsymbol{h} \in H^*, \boldsymbol{h}' \in H^*}$. C is invertible as it is upper triangular with non-zero diagonal entries. We are now ready to prove Theorem A.1.

Theorem A.1. *For any two distributions* Q, R *on* $(0,1)^K$, $\tilde{\boldsymbol{\pi}}_Q = \tilde{\boldsymbol{\pi}}_R$ *is equivalent to* $\boldsymbol{m}_Q = A\boldsymbol{m}_R$ *for some* $A > 0$.

Proof. $\tilde{\boldsymbol{\pi}}_Q = \tilde{\boldsymbol{\pi}}_R$ is equivalent to $\boldsymbol{\pi}_Q^*/(1-\pi_{Q,0}) = \boldsymbol{\pi}_R^*/(1-\pi_{R,0})$. Rearranging terms we have that $\boldsymbol{\pi}_Q^* = \boldsymbol{\pi}_R^*(1-\pi_{Q,0})/(1-\pi_{R,0})$, and thus $\boldsymbol{\pi}_Q^* = A\boldsymbol{\pi}_R^*$, where $A = (1-\pi_{Q,0})/(1-\pi_{R,0}) > 0$. Using Lemma A.1, this is equivalent

to $C\boldsymbol{m}_Q = AC\boldsymbol{m}_R$, and thus $\boldsymbol{m}_Q = A\boldsymbol{m}_R$ due to the invertibility of C. □

The immediate consequence of Theorem A.1 is that to verify conditional identifiability of an M_{th} model $\mathcal{P}_{\Omega_{\mathcal{Q}}}$, one can demonstrate that if $\boldsymbol{m}_Q = A\boldsymbol{m}_R$ for some $Q, R \in \mathcal{Q}$, then $\pi_{Q,0} = \pi_{R,0}$. We use this mechanism in the next section to characterize when latent class models (LCMs) are conditionally identifiable.

A.3 Conditional Identifiability of Latent Class Models

We denote the family of mixing distributions corresponding to LCMs with J classes by $\mathcal{Q}_J = \{Q = \sum_{j=1}^{J} \nu_{Q,j} \prod_{k=1}^{K} \delta_{q_{Q,jk}} \mid \nu_{Q,j} \geq 0, \sum_{j=1}^{J} \nu_{Q,j} = 1, q_{Q,jk} \in (0,1)^K\}$, so that $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ is equivalent to $\mathcal{P}_{\Omega_{LCM,J}}$ from the main text. To provide necessary and sufficient conditions for $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ to be conditionally identifiable, we restrict the family of mixing distributions to $\mathcal{Q}_J = \{Q = \sum_{j=1}^{J} \nu_{Q,j} \prod_{k=1}^{K} \delta_{q_{Q,jk}} \mid \nu_{Q,j} \geq 0, \sum_{j=1}^{J} \nu_{Q,j} = 1, q_{Q,jk} \in (0,1)^K, q_{Q,jk} \neq q_{Q,j'k} \text{ for } j \neq j'\}$. This restriction makes the mild assumption that each class' sampling probabilities are distinct, which simplifies the proof of Theorem A.2. Loosening this restriction could only make the conditions on J for $\mathcal{Q}_J$ to be identifiable stricter, and thus the conclusions we reach in Section A.6 would still stand for families where this restriction is violated.

There are $J(K+1)-1$ parameters in $\mathcal{Q}_J$, thus when $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ is conditionally identifiable, J satisfies $J(K+1)-1 \leq 2^K - 2$, as the observed cell probabilities, $\tilde{\boldsymbol{\pi}}_Q$, are $2^K - 2$ dimensional. However, we now prove that J must satisfy a stricter condition for $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ to be conditionally identifiable. In Section A.6 we discuss some limitations of this result.

Theorem A.2. *$\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ is conditionally identifiable iff $2J \leq K$.*

Proof. We will first show that if $2J \leq K$, then $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ is conditionally identifiable. The proof of this direction is similar in spirit to the proofs of Theorem 2 in Holzmann et al. (2006) and Theorem 1 in Pezzott et al. (2019), which were both concerned with characterizing the identifiability of the M_h analogue of $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$. Assume $2J \leq K$, and let $Q, R \in \mathcal{Q}_J$ such that $\boldsymbol{m}_Q = A\boldsymbol{m}_R$ for some $A > 0$, so that we have the following system of equations:

$$\sum_{j=1}^{J} \nu_{Q,j} \prod_{k=1}^{K} q_{Q,jk}^{h_k} - A \sum_{j=1}^{J} \nu_{R,j} \prod_{k=1}^{K} q_{R,jk}^{h_k} = 0 \quad (\boldsymbol{h} \in H^*). \tag{A.2}$$

Let $\mathcal{I}_Q = \{j \mid q_{Q,j} \notin (q_{R,1}, \ldots, q_{R,J})\}$ and $\mathcal{I}_R = \{j \mid q_{R,j} \notin (q_{Q,1}, \ldots, q_{Q,J})\}$, where $q_{Q,j} = (q_{Q,j1}, \ldots, q_{Q,jK})$ and $q_{R,j} = (q_{R,j1}, \ldots, q_{R,jK})$. We can then rewrite (A.2) as

$$\sum_{j=1}^{J} y_j \prod_{k=1}^{K} q_{Q,jk}^{h_k} - A \sum_{i \in \mathcal{I}_R}^{J} \nu_{R,j} \prod_{k=1}^{K} q_{R,jk}^{h_k} = 0 \quad (\boldsymbol{h} \in H^*), \tag{A.3}$$

where $y_j = \nu_{Q,j}$ if $j \in \mathcal{I}_Q$ and $y_j = \nu_{Q,j} - A\nu_{R,j'}$ for some $j' \in \{1, \ldots, J\} \setminus \mathcal{I}_R$ otherwise. Letting $m = |\mathcal{I}_R| = |\mathcal{I}_Q|$ and labelling the elements of $\mathcal{I}_R$ as $i_1, \ldots, i_m$, the system of equations in (A.3) can be written in matrix form as $\Lambda \boldsymbol{y} = 0$, where

$$\Lambda = \begin{pmatrix} q_{Q,1K} & \cdots & q_{Q,JK} & q_{R,i_1K} & \cdots & q_{R,i_mK} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \prod_{k=1}^{K} q_{Q,1k}^{h_k} & \cdots & \prod_{k=1}^{K} q_{Q,Jk}^{h_k} & \prod_{k=1}^{K} q_{R,i_1k}^{h_k} & \cdots & \prod_{k=1}^{K} q_{R,i_mk}^{h_k} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \prod_{k=1}^{K} q_{Q,1k} & \cdots & \prod_{k=1}^{K} q_{Q,Jk} & \prod_{k=1}^{K} q_{R,i_1k} & \cdots & \prod_{k=1}^{K} q_{R,i_mk} \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_J \\ -A\nu_{R,i_1} \\ \vdots \\ -A\nu_{R,i_m} \end{pmatrix},$$

and the rows of Λ are indexed by $\boldsymbol{h} \in H^*$. In Section A.4, we prove that Λ is full rank, and thus $\boldsymbol{y} = 0$, for any $m \in \{0, \ldots, J\}$. The proof of this direction concludes by examining three possible cases.

Case 1. *Suppose $m = 0$, i.e. for each $j \in \{1, \ldots, J\}$, there exists some $j' \in \{1, \ldots, J\}$ such that $q_{Q,j} = q_{R,j'}$ and $\nu_{Q,j} = A\nu_{R,j'}$. As $\sum_{j=1}^{J} \nu_{Q,j} = \sum_{j=1}^{J} \nu_{R,j} = 1$, this implies that $A = 1$ and thus $\pi_{Q,0} = \pi_{R,0}$.*

Case 2. *Suppose $m \in \{1, \ldots, J-1\}$, i.e. for each $j \in \{1, \ldots, J\} \setminus \mathcal{I}_Q$, there exists some $j' \in \{1, \ldots, J\} \setminus \mathcal{I}_R$ such that $q_{Q,j} = q_{R,j'}$ and $\nu_{Q,j} = A\nu_{R,j'}$. Further, for each $j \in \mathcal{I}_Q$ and $j' \in \mathcal{I}_R$ $\nu_{Q,j} = \nu_{R,j'} = 0$. We can thus ignore the classes $j \in \mathcal{I}_Q$ and $j' \in \mathcal{I}_R$. As $\sum_{j=1}^{J} \nu_{Q,j} = \sum_{j=1}^{J} \nu_{R,j} = 1$, this implies that $A = 1$ and thus $\pi_{Q,0} = \pi_{R,0}$.*

Case 3. *Suppose $m = J$, i.e. for each $j \in \{1, \ldots, J\}$, there exists no $j' \in \{1, \ldots, J\}$ such that $q_{Q,j} = q_{R,j'}$. Then $\nu_{Q,j} = \nu_{R,j} = 0$ for $j \in \{1, \ldots, J\}$, which is a contradiction.*

We will now show that if $2J > K$, then $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ is not conditionally identifiable. To do so we will provide explicit $Q, R \in \mathcal{Q}_J$ such that $\pi_{Q,0} \neq \pi_{R,0}$, but $\boldsymbol{m}_Q = A\boldsymbol{m}_R$ for $A > 0$. This counterexample is modified from Tahmasebi et al. (2018), who studied identifiability of families of LCMs outside of the multiple-systems estimation

context where n_0 is observed. Choose J such that $2J > K$. For $j \in \{1, \dots, J\}$, let $\nu_{Q,j} = \binom{2J}{2j}/(2^{2J-1} - 1)$ and $\nu_{R,j} = \binom{2J}{2j-1}/(2^{2J-1})$. For $j \in \{1, \dots, J\}$ and $k \in \{1, \dots, K\}$, let $q_{Q,jk} = \alpha(2j)$ and $q_{R,jk} = \alpha(2j-1)$ where $0 < \alpha < 1/(2J)$. We thus have that $Q, R \in \mathcal{Q}_J$, where clearly $Q \neq R$. In Section A.5 we prove that for these choices of Q, R, $\boldsymbol{m}_Q = A\boldsymbol{m}_R$ for $A > 0$ such that $A \neq 1$, and thus $\pi_{Q,0} \neq \pi_{R,0}$. □

A.4 Proof that Λ is Full Rank

We will prove that Λ is full rank for any $m \in \{0, \dots, J\}$ by proving a stronger result. Recall that $K \geq 2$ and let $x_{\ell k} \in (0, 1)$ for $\ell \in \{1, \dots, K\}$ and $k \in \{1, \dots, K\}$, such that $x_{\ell k} \neq x_{\ell k'}$ for $k \neq k'$. Let

$$X^K = \begin{pmatrix} x_{1K} & \cdots & x_{KK} \\ \vdots & \ddots & \vdots \\ \prod_{k=1}^K x_{1k}^{h_k} & \cdots & \prod_{k=1}^K x_{Kk}^{h_k} \\ \vdots & \ddots & \vdots \\ \prod_{k=1}^K x_{1k} & \cdots & \prod_{k=1}^K x_{Kk} \end{pmatrix},$$

where the rows of X^K are indexed by $\boldsymbol{h} \in H^*$. We will show that X^K is full rank by induction on K. This implies that Λ is full rank, as $J + m \leq 2J \leq K$ by assumption for any $m \in \{0, \dots, J\}$.

For the base case when $K = 2$, verifying X^2 is full rank is straightforward. Assume that X^{K-1} is full rank. Let $\boldsymbol{v} \in \mathbb{R}^{K \times 1}$ be such that $X^K \boldsymbol{v} = 0$. For each $\boldsymbol{h} \in \{\boldsymbol{h}' \in H^* \mid h'_K = 0\}$ we have that $v_K \prod_{k=1}^{K-1} x_{Kk}^{h_k} = -\sum_{\ell=1}^{K-1} v_\ell \prod_{k=1}^{K-1} x_{\ell k}^{h_k}$, which implies that $\sum_{\ell=1}^{K-1} v_\ell (x_{\ell K} - x_{KK}) \prod_{k=1}^{K-1} x_{\ell k}^{h_k} = 0$. For $\ell \in \{1, \dots, K-1\}$, let $v'_\ell = v_\ell(x_{\ell K} - x_{KK})$ and $\boldsymbol{v}' = (v'_1, \dots, v'_{K-1})$. This leads to the system of equations $X^{K-1}\boldsymbol{v}' = 0$. By the inductive assumption, $\boldsymbol{v}' = 0$. Since $x_{\ell K} \neq x_{KK}$ for $\ell \in \{1, \dots, K-1\}$, we have that $v_\ell = 0$ for $\ell \in \{1, \dots, K-1\}$, and thus $v_K = 0$.

A.5 Proof of Counterexample

We will now prove that $m_{Q,\boldsymbol{h}} = A m_{R,\boldsymbol{h}}$ for all $\boldsymbol{h} \in H^*$, where $A = (2^{2J-1})/(2^{2J-1} - 1) \neq 1$. Define the function $h(x) = (1 - e^{\alpha x})^{2J} = \sum_{i=0}^{2J} \binom{2J}{i} (-1)^i e^{\alpha i x}$. For $t \in \{1, \dots, K\}$, we can differentiate the series representation of h to

find that $h^{(t)}(x) = \sum_{i=0}^{2J} \binom{2J}{i}(-1)^i(\alpha i)^t e^{\alpha i x}$ and thus $h^{(t)}(x)|_{x=0} = \sum_{i=0}^{2J} \binom{2J}{i}(-1)^i(\alpha i)^t = \sum_{i=1}^{2J} \binom{2J}{i}(-1)^i(\alpha i)^t$. We can alternatively differentiate the non-series representation of h using the fact that $t \leq K < 2J$ and the chain rule for higher order derivatives to find that $h^{(t)}(x)|_{x=0} = 0$. Let $\boldsymbol{h} \in H^*$ and $t = \sum_{k=1}^{K} h_k \in \{1, \ldots, K\}$. The desired result follows as

$$
\begin{aligned}
m_{Q,\boldsymbol{h}} - A m_{R,\boldsymbol{h}} &= \sum_{j=1}^{J} \nu_{Q,j} \prod_{k=1}^{K} q_{Q,jk}^{h_k} - A \sum_{j=1}^{J} \nu_{R,j} \prod_{k=1}^{K} q_{R,jk}^{h_k} \\
&= \sum_{j=1}^{J} \binom{2J}{2j} (2^{2J-1} - 1)^{-1} \prod_{k=1}^{K} \{\alpha(2j)\}^{h_k} - A \sum_{j=1}^{J} \binom{2J}{2j-1} (2^{2J-1})^{-1} \prod_{k=1}^{K} \{\alpha(2j-1)\}^{h_k} \\
&= (2^{2J-1} - 1)^{-1} \sum_{i=1}^{2J} \binom{2J}{i} (-1)^i (\alpha i)^t = (2^{2J-1} - 1)^{-1} \{h^{(t)}(x)|_{x=0}\} = 0.
\end{aligned}
$$

A.6 Limitations of Theorem A.2

Theorem A.2 shows that $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ is not conditionally identifiable if $2J > K$ by counterexample, by demonstrating two mixing distributions $Q, R \in \mathcal{Q}_J$ where $\tilde{\boldsymbol{\pi}}_Q = \tilde{\boldsymbol{\pi}}_R$ but $\pi_{Q,0} \neq \pi_{R,0}$. Within each latent class of Q and R, the sampling probabilities were the same, meaning Q and R can be seen as mixing distributions of an M_h model. It would be interesting in future work to see whether further restrictions on $\Omega_{\mathcal{Q}_J}$, for example restrictions not allowing the sampling probabilities within latent classes to be equal, lead to different results concerning conditional identifiability. Another interesting route would be to see whether results concerning *generic identifiability* of latent class models (Allman et al., 2009) could be applied to the multiple-systems estimation setting.

However, this does not mean Theorem A.2 is not a practically useful result. Theorem A.2 provides assumptions under which which we have formal statistical guarantees for when we can estimate the parameters in $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$: the parameters of $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ can be consistently estimated if $2J \leq K$. When $2J > K$ we currently have no such guarantees. In Web Appendix D we demonstrate this reality across a variety of simulation studies.

B Web Appendix B: Computation for Conditionally Identified Models

The purpose of this appendix is to provide details of how computation for conditionally identified models can be carried out in both frequentist and Bayesian frameworks using existing software. Recall from Sections 2.3 and 2.4

of the main text that the complete-data distribution can be written as

$$p(\boldsymbol{n}, n_0 \mid N, \boldsymbol{\pi}) = N! \prod_{\boldsymbol{h} \in H} \frac{\pi_{\boldsymbol{h}}^{n_{\boldsymbol{h}}}}{n_{\boldsymbol{h}}!} = L_1(N, \pi_0 \mid n) L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}), \tag{B.1}$$

with $L_1(N, \pi_0 \mid n) = \binom{N}{n} \pi_0^{N-n} (1-\pi_0)^n$ and $L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) = n! \prod_{\boldsymbol{h} \in H^*} \tilde{\pi}_{\boldsymbol{h}}^{n_{\boldsymbol{h}}} / n_{\boldsymbol{h}}!$, and that conditionally identified models have parameter spaces of the form $\Omega = \{N, \pi_0, \tilde{\boldsymbol{\pi}} \mid N \in \mathbb{N}, \pi_0 = \mathcal{T}(\tilde{\boldsymbol{\pi}}), \tilde{\boldsymbol{\pi}} \in \tilde{S}\}$.

B.1 Computation for Frequentist Multiple-Systems Estimation

In this section we will first describe an approach for frequentist inference in general conditionally identified models, followed by the specific cases of models using the NHOI and the K'-list marginal NHOI identifying assumptions.

B.1.1 Conditionally Identified Models in General

Suppose that we are using a conditionally identified model with parameter space $\Omega = \{N, \pi_0, \tilde{\boldsymbol{\pi}} \mid N \in \mathbb{N}, \pi_0 = \mathcal{T}(\tilde{\boldsymbol{\pi}}), \tilde{\boldsymbol{\pi}} \in \tilde{S}\}$. Frequentist inference for this general conditionally identified model will follow from the conditional maximum likelihood approach outlined in Sanathanan (1972) and Fienberg (1972). In particular, this approach can be summarized in two steps:

1. Estimate the observed cell probabilities $\tilde{\boldsymbol{\pi}}$ by maximizing the conditional likelihood over the set of possible observed cell probabilities $\tilde{S}$:

$$\hat{\tilde{\boldsymbol{\pi}}} = \arg\max_{\tilde{\boldsymbol{\pi}} \in \tilde{S}} L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}).$$

2. Estimate the population size N by maximizing the binomial likelihood for n conditional on the estimate of the observed cell probabilities, $\hat{\tilde{\boldsymbol{\pi}}}$:

$$\hat{N}(\hat{\tilde{\boldsymbol{\pi}}}) = \arg\max_{N \in \mathbb{N}} L_1(N, \mathcal{T}(\hat{\tilde{\boldsymbol{\pi}}}) \mid n) = \left\lfloor \frac{n}{1 - \mathcal{T}(\hat{\tilde{\boldsymbol{\pi}}})} \right\rfloor,$$

where $\lfloor \cdot \rfloor$ is the floor function. We will ignore the rounding and write the estimator of N as $\hat{N}(\hat{\tilde{\boldsymbol{\pi}}}) = n / \{1 - \mathcal{T}(\hat{\tilde{\boldsymbol{\pi}}})\}$. This is well known as the Horvitz-Thompson estimator (Horvitz and Thompson, 1952).

We note here that $\hat{\tilde{\boldsymbol{\pi}}}$, and thus $\hat{N}(\hat{\tilde{\boldsymbol{\pi}}})$, may not exist in general, depending on the set of possible observed cell probabilities $\tilde{S}$. The sample proportions $\{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h} \in H^*}$ maximize the conditional likelihood over $\mathbb{S}^{2^K - 2}$, so if the

sample proportions lie in $\tilde{S}$, then they maximize the conditional likelihood over $\tilde{S}$. If the sample proportions do not lie in $\tilde{S}$, care must be taken to make sure that $\hat{\boldsymbol{\pi}}$ exists.

For the rest of this section we will assume that the model is correctly specified, and $\hat{\boldsymbol{\pi}}$ exists. Let $\tilde{\boldsymbol{\pi}}^*$ denote the true observed cell probabilities. Suppose it is true, for an estimator $\hat{\boldsymbol{\pi}}$ of $\tilde{\boldsymbol{\pi}}^*$, that $\sqrt{n}(\hat{\boldsymbol{\pi}} - \tilde{\boldsymbol{\pi}}^*) \mid n \rightarrow_d \textsc{Normal}(0, \Sigma(\tilde{\boldsymbol{\pi}}^*))$, where $\rightarrow_d$ denotes convergence in distribution and we are conditioning on n (i.e. ignoring binomial variation in n). For example, when the sample proportions $\{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h} \in H^*}$ lie within $\tilde{S}$, we have that $\hat{\boldsymbol{\pi}} = \{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h} \in H^*}$ and $\Sigma(\tilde{\boldsymbol{\pi}}^*) = \text{diag}(\tilde{\boldsymbol{\pi}}^*) - \tilde{\boldsymbol{\pi}}^*(\tilde{\boldsymbol{\pi}}^*)^T$ (see e.g. chapter 14 of Agresti, 2003). For $\tilde{\boldsymbol{\pi}} \in \tilde{S}$, let $f(\tilde{\boldsymbol{\pi}}) = 1/(1 - \mathcal{T}(\tilde{\boldsymbol{\pi}}))$. From the delta method, it follows that $\sqrt{n}(f(\hat{\boldsymbol{\pi}}) - f(\tilde{\boldsymbol{\pi}}^*)) \mid n \rightarrow_d \textsc{Normal}(0, (\nabla f(\tilde{\boldsymbol{\pi}}^*))^T \Sigma(\tilde{\boldsymbol{\pi}}^*) \nabla f(\tilde{\boldsymbol{\pi}}^*))$. Thus for large n, $nf(\hat{\boldsymbol{\pi}}) = \hat{N}(\hat{\boldsymbol{\pi}}) \approx \textsc{Normal}(nf(\tilde{\boldsymbol{\pi}}^*), n(\nabla f(\tilde{\boldsymbol{\pi}}^*))^T \Sigma(\tilde{\boldsymbol{\pi}}^*) \nabla f(\tilde{\boldsymbol{\pi}}^*))$. We can then substitute our estimate $\hat{\boldsymbol{\pi}}$ of the observed cell probabilities for $\tilde{\boldsymbol{\pi}}^*$, and use this large sample approximation to construct 95% confidence intervals for N of the form $\hat{N}(\hat{\boldsymbol{\pi}}) \pm 1.96 * \sqrt{n(\nabla f(\hat{\boldsymbol{\pi}}))^T \Sigma(\hat{\boldsymbol{\pi}}) \nabla f(\hat{\boldsymbol{\pi}})}$. The term $(\nabla f(\hat{\boldsymbol{\pi}}))^T \Sigma(\hat{\boldsymbol{\pi}}) \nabla f(\hat{\boldsymbol{\pi}})$ can be calculated automatically using e.g. the `delta.method` function in the `R` package `msm` (Jackson, 2011).

The confidence interval construction in the last paragraph conditions on n, and thus does not incorporate the binomial variation of n. Let N^* denote the true population size. For $\tilde{\boldsymbol{\pi}} \in \tilde{S}$, let $g(\tilde{\boldsymbol{\pi}}) = \mathcal{T}(\tilde{\boldsymbol{\pi}})/(1 - \mathcal{T}(\tilde{\boldsymbol{\pi}}))$. Following Fienberg (1972), unconditional of n we have that $(N^*)^{-1/2}(\hat{N}(\hat{\boldsymbol{\pi}}) - N^*) \rightarrow_d \textsc{Normal}(0, g(\tilde{\boldsymbol{\pi}}^*) + (1 - \mathcal{T}(\tilde{\boldsymbol{\pi}}^*))(\nabla g(\tilde{\boldsymbol{\pi}}^*))^T \Sigma(\tilde{\boldsymbol{\pi}}^*) \nabla g(\tilde{\boldsymbol{\pi}}^*))$. Thus for large N^*, $\hat{N}(\hat{\boldsymbol{\pi}}) \approx \textsc{Normal}(N^*, N^* g(\tilde{\boldsymbol{\pi}}^*) + N^*(1 - \mathcal{T}(\tilde{\boldsymbol{\pi}}^*))(\nabla g(\tilde{\boldsymbol{\pi}}^*))^T \Sigma(\tilde{\boldsymbol{\pi}}^*) \nabla g(\tilde{\boldsymbol{\pi}}^*))$. We can then substitute our estimate $\hat{\boldsymbol{\pi}}$ of the observed cell probabilities for $\tilde{\boldsymbol{\pi}}^*$ and our estimate $\hat{N}(\hat{\boldsymbol{\pi}})$ of the population size for N^*, and use this large sample approximation to construct 95% confidence intervals for N of the form $\hat{N}(\hat{\boldsymbol{\pi}}) \pm 1.96 * \sqrt{\hat{N}(\hat{\boldsymbol{\pi}}) g(\hat{\boldsymbol{\pi}}) + n(\nabla g(\hat{\boldsymbol{\pi}}))^T \Sigma(\hat{\boldsymbol{\pi}}) \nabla g(\hat{\boldsymbol{\pi}})}$. Again, the term $(\nabla g(\hat{\boldsymbol{\pi}}))^T \Sigma(\hat{\boldsymbol{\pi}}) \nabla g(\hat{\boldsymbol{\pi}})$ can be calculated automatically using e.g. the `delta.method` function in the `R` package `msm` (Jackson, 2011).

B.1.2 Computation for the NHOI and K'-List Marginal NHOI Identifying Assumptions

In this section we will focus on frequentist inference in the specific cases of models using the NHOI and the K'-list marginal NHOI identifying assumptions. While one could construct estimators and confidence intervals for N, under these assumptions, by hand using the results from the previous section, software is already available which accomplishes these tasks.

NHOI Identifying Assumption

For the NHOI identifying assumption, there are many `R` packages which produce estimates and confidence intervals for the population size under this assumption. For example, in our Kosovo application we use the `Rcapture` package Baillargeon et al. (2007). The function `closedpMS.t` produces estimates and standard errors for the population size under all hierarchical log-linear models, including the saturated log-linear model $\mathcal{P}_{\Omega_{LL}}$. These can then be used to construct confidence intervals for the population size.

K'-List Marginal NHOI Identifying Assumption

Recall from Section 4.3 of the main text that the K'-list marginal NHOI identifying assumption restricts the observed cell probabilities to lie in $\tilde{S} = \{\tilde{\boldsymbol{\pi}} \in \mathbb{S}^{2^K-2} \mid \tilde{\Pi}_{odd,+}/(\tilde{\Pi}_{even,+}\tilde{\pi}_{0+}) > 1\}$. Thus there are two cases to consider when fitting a model in the frequentist framework using the K'-list marginal NHOI identifying assumption:

1. The sample proportions $\{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h}\in H^*}$ lie within $\tilde{S} = \{\tilde{\boldsymbol{\pi}} \in \mathbb{S}^{2^K-2} \mid \tilde{\Pi}_{odd,+}/(\tilde{\Pi}_{even,+}\tilde{\pi}_{0+}) > 1\}$.
2. The sample proportions $\{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h}\in H^*}$ do not lie within $\tilde{S}$.

There is a simple way to verify for a given data set, which case one is in. Consider the restricted data set from just the first K' lists. In particular, using notation from Section 4.3 of the main text, $\{n^\dagger_{\boldsymbol{g}}\}_{\boldsymbol{g}\in G^*}$ is the restricted data, where $n^\dagger_{\boldsymbol{g}} = \sum_{\boldsymbol{h}\in H^*} n_{\boldsymbol{h}} I\{(h_1, \cdots, h_{K'}) = \boldsymbol{g}\}$, and the restricted sample size is $n^\dagger = \sum_{\boldsymbol{g}\in G^*} n^\dagger_{\boldsymbol{g}}$. Using this restricted data set of K' lists, one could compute the frequentist population size estimator under the NHOI assumption (for K' lists), using standard software (e.g., the `Rcapture` package as just described). Call this estimate $\hat{N}^\dagger$. Then the sample proportions $\{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h}\in H^*}$ lie within $\tilde{S}$ as long as $\hat{N}^\dagger > n$.

Suppose we are in the second case, i.e. the sample proportions $\{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h}\in H^*}$ do not lie in $\tilde{S}$. In this case, $\hat{\boldsymbol{\pi}}$ may not exist. One needs to verify that $\hat{\boldsymbol{\pi}}$ exists, and if it does, compute it and derive its asymptotic distribution to compute confidence intervals for N as described in Appendix B.1.1. This could potentially be quite difficult technically, so we recommend if one truly believes that the K'-list marginal NHOI identifying assumption holds in this case, that they use a Bayesian estimator as described in Appendix B.2.

Suppose now we are in the first case, i.e. the sample proportions $\{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h}\in H^*}$ lie in $\tilde{S}$. Then $\hat{\boldsymbol{\pi}} = \{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h}\in H^*}$, and thus we could then follow the details at the end of Appendix B.1.1 to arrive at a confidence interval for N.

However, we want to take advantage of existing software in order to compute estimates and confidence intervals for N. The following theorem accomplishes this task:

Theorem B.1. *Let $\hat{N}$ denote the population size estimator under the K'-list marginal NHOI identifying assumption using the full K list data set. Let $\hat{N}^{\dagger}$ denote the population size estimator when restricting to data from just the first K' lists and using the NHOI assumption for K' lists. If the sample proportions $\{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h}\in H^*}$ lie in $\tilde{S} = \{\tilde{\boldsymbol{\pi}} \in \mathbb{S}^{2^K-2} \mid \tilde{\Pi}_{odd,+}/(\tilde{\Pi}_{even,+}\tilde{\pi}_{0+}) > 1\}$, then $\hat{N} = \hat{N}^{\dagger}$.*

We prove Theorem B.1 in Appendix B.1.3. Theorem B.1 tells us that if we want to calculate estimates and confidence intervals for the population size under the K'-list marginal NHOI identifying assumption, we can accomplish this by restricting the data set to K' lists, and calculating estimates and confidence intervals for the population size estimate for just these K' lists under the NHOI assumption for K' lists. This can be accomplished using the function `closedpMS.t` in the `Rcapture` package.

Sensitivity Analyses

`Rcapture` does not support sensitivity analyses that examine the impact of the NHOI or K'-list marginal NHOI identifying assumptions, as described in Section 4.2 and 4.3 of the main text. However, it is straightforward to use the `glm` function in `R` to perform these sensitivity analyses, which is what `Rcapture` uses under the hood. In the code accompanying this manuscript, available at `github.com/aleshing/central-role-of-identifying-assumptions`, we provide a function which performs these sensitivity analyses.

B.1.3 Proof of Theorem B.1

Proof. Suppose we have data from K lists, $\{n_{\boldsymbol{h}}\}_{\boldsymbol{h}\in H^*}$, with observed sample size n, and we are using the K'-list marginal NHOI assumption, for $1 < K' < K$. For this proof, denote the sample proportions by $\tilde{\boldsymbol{\pi}} = \{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h}\in H^*}$.

We start by restating some notation from Section 4.3 of the main paper. Let $G = \{0,1\}^{K'}$ index the marginal $2^{K'}$ contingency table for the first K' lists and $G^* = G \setminus \{0\}^{K'}$. Let $\tilde{\pi}_{\boldsymbol{g}+} = \sum_{\boldsymbol{h}\in H^*} \tilde{\pi}_{\boldsymbol{h}} I\{(h_1,\cdots,h_{K'}) = \boldsymbol{g}\}$ and $\tilde{\pi}_{0+} = \sum_{\boldsymbol{h}\in H^*} \tilde{\pi}_{\boldsymbol{h}} I\{(h_1,\cdots,h_{K'}) = (0,\cdots,0)\}$. The K'-lists marginal NHOI assumption corresponds to the explicit identifying assumption $\mathcal{T}(\tilde{\boldsymbol{\pi}}) = (\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \tilde{\pi}_{0+})/(1 + \tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \tilde{\pi}_{0+})$, where $\tilde{\Pi}_{odd,+} =$

$\prod_{\boldsymbol{g}\in G^*} \tilde{\pi}_{\boldsymbol{g}+}^{I_{odd}(\boldsymbol{g})}$ and $\tilde{\Pi}_{even,+} = \prod_{\boldsymbol{g}\in G^*} \tilde{\pi}_{\boldsymbol{g}+}^{I_{even}(\boldsymbol{g})}$. $\mathcal{T}(\tilde{\boldsymbol{\pi}}) \in (0,1)$ since we assume that $\tilde{\Pi}_{odd,+}/(\tilde{\Pi}_{even,+}\tilde{\pi}_{0+}) > 1$.

Now we introduce some new notation. Suppose we are restricted to just the data from the first K' lists. Let $\{n_{\boldsymbol{g}}^\dagger\}_{\boldsymbol{g}\in G^*}$ denote the restricted data, so that $n_{\boldsymbol{g}}^\dagger = \sum_{\boldsymbol{h}\in H^*} n_{\boldsymbol{h}} I\{(h_1,\cdots,h_{K'}) = \boldsymbol{g}\}$, and the restricted sample size is $n^\dagger$. Denote the restricted sample proportions by $\tilde{\boldsymbol{\pi}}^\dagger = \{n_{\boldsymbol{g}}^\dagger/n^\dagger\}_{\boldsymbol{g}\in G^*}$. Using this restricted K' list data set, the NHOI assumption corresponds to the explicit identifying assumption $\mathcal{T}^\dagger(\tilde{\boldsymbol{\pi}}^\dagger) = (\tilde{\Pi}_{odd}^\dagger/\tilde{\Pi}_{even}^\dagger)/(1+\tilde{\Pi}_{odd}^\dagger/\tilde{\Pi}_{even}^\dagger)$, where $\tilde{\Pi}_{odd}^\dagger = \prod_{\boldsymbol{g}\in G^*}(\tilde{\pi}_{\boldsymbol{g}}^\dagger)^{I_{odd}(\boldsymbol{g})}$ and $\tilde{\Pi}_{even}^\dagger = \prod_{\boldsymbol{g}\in G^*}(\tilde{\pi}_{\boldsymbol{g}}^\dagger)^{I_{even}(\boldsymbol{g})}$.

In a frequentist framework, the population size estimate using the K'-list marginal NHOI assumption when the estimated observed cell probabilities are $\tilde{\boldsymbol{\pi}}$ is

$$\hat{N} = \frac{n}{1-\frac{\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+}-\tilde{\pi}_{0+}}{1+\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+}-\tilde{\pi}_{0+}}} = n\left[1+\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+}-\tilde{\pi}_{0+}\right].$$

Similarly, in a frequentist framework the population size estimate using the NHOI assumption with the restricted K' list data set is

$$\hat{N}^\dagger = \frac{n^\dagger}{1-\frac{\tilde{\Pi}_{odd}^\dagger/\tilde{\Pi}_{even}^\dagger}{1+\tilde{\Pi}_{odd}^\dagger/\tilde{\Pi}_{even}^\dagger}} = n^\dagger\left[1+\tilde{\Pi}_{odd}^\dagger/\tilde{\Pi}_{even}^\dagger\right].$$

Our task is to prove that $\hat{N} = \hat{N}^\dagger$.

We list here two useful facts that can be verified through simple algebra:

1. $n^\dagger = n - n\tilde{\pi}_{0+} = n[1-\tilde{\pi}_{0+}]$.

2. $\tilde{\pi}_{\boldsymbol{g}} = \tilde{\pi}_{\boldsymbol{g}}^\dagger[1-\tilde{\pi}_{0+}]$.

Using the first fact, we can rewrite $\hat{N}^\dagger$:

$$\begin{aligned}\hat{N}^\dagger &= n^\dagger\left[1+\tilde{\Pi}_{odd}^\dagger/\tilde{\Pi}_{even}^\dagger\right]\\ &= n[1-\tilde{\pi}_{0+}]\left[1+\tilde{\Pi}_{odd}^\dagger/\tilde{\Pi}_{even}^\dagger\right]\\ &= n\left[1+(1-\tilde{\pi}_{0+})(\tilde{\Pi}_{odd}^\dagger/\tilde{\Pi}_{even}^\dagger)-\tilde{\pi}_{0+}\right].\end{aligned}$$

Thus if we can show that $(1-\tilde{\pi}_{0+})(\tilde{\Pi}_{odd}^\dagger/\tilde{\Pi}_{even}^\dagger) = \tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+}$, the proof is complete. Using the second fact,

we can rewrite $(1-\tilde{\pi}_{0+})(\tilde{\Pi}^{\dagger}_{odd}/\tilde{\Pi}^{\dagger}_{even})$:

$$\begin{aligned}(1-\tilde{\pi}_{0+})(\tilde{\Pi}^{\dagger}_{odd}/\tilde{\Pi}^{\dagger}_{even}) &= (1-\tilde{\pi}_{0+})\left[\frac{\prod_{\boldsymbol{g}\in G^*}(\tilde{\pi}^{\dagger}_{\boldsymbol{g}})^{I_{odd}(\boldsymbol{g})}}{\prod_{\boldsymbol{g}\in G^*}(\tilde{\pi}^{\dagger}_{\boldsymbol{g}})^{I_{even}(\boldsymbol{g})}}\right] \\ &= (1-\tilde{\pi}_{0+})\left[\frac{1-\tilde{\pi}_{0+}}{1-\tilde{\pi}_{0+}}\right]^{2^{K'-1}}\left[\frac{\prod_{\boldsymbol{g}\in G^*}(\tilde{\pi}^{\dagger}_{\boldsymbol{g}})^{I_{odd}(\boldsymbol{g})}}{\prod_{\boldsymbol{g}\in G^*}(\tilde{\pi}^{\dagger}_{\boldsymbol{g}})^{I_{even}(\boldsymbol{g})}}\right] \\ &= \left[\frac{\prod_{\boldsymbol{g}\in G^*}\tilde{\pi}^{I_{odd}(\boldsymbol{g})}_{\boldsymbol{g}+}}{\prod_{\boldsymbol{g}\in G^*}\tilde{\pi}^{I_{even}(\boldsymbol{g})}_{\boldsymbol{g}+}}\right] = \tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+}.\end{aligned}$$

□

B.2 Computation for Bayesian Multiple-Systems Estimation

In this section we will describe a computational approach for Bayesian inference in general conditionally identified models, that allows any prior for the population size, N, and any prior for the observed cell probabilities, $\tilde{\boldsymbol{\pi}}$. Various sensitivity analyses are facilitated from this approach. We further give some guidance to specification of the prior for N.

B.2.1 Bayesian Multiple-Systems Estimation

Suppose that we are using a conditionally identified model with parameter space $\Omega = \{N, \pi_0, \tilde{\boldsymbol{\pi}} \mid N \in \mathbb{N}, \pi_0 = \mathcal{T}(\tilde{\boldsymbol{\pi}}), \tilde{\boldsymbol{\pi}} \in \tilde{S}\}$, and we have specified independent prior distributions for N and $\tilde{\boldsymbol{\pi}}$, with densities $p(N)$ and $p(\tilde{\boldsymbol{\pi}})$. In this section, and the following two sections, we will let $p(\cdot)$ denote a density of a given random variable. The joint posterior of N and $\tilde{\boldsymbol{\pi}}$ can be written as $p(N, \tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) \propto L_1(N, \mathcal{T}(\tilde{\boldsymbol{\pi}}) \mid n) L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) p(N) p(\tilde{\boldsymbol{\pi}}) I(\tilde{\boldsymbol{\pi}} \in \tilde{S})$. The marginal posteriors of $\tilde{\boldsymbol{\pi}}$ and N can be written as

$$p(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) \propto p(n \mid \mathcal{T}(\tilde{\boldsymbol{\pi}})) L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) p(\tilde{\boldsymbol{\pi}}) I(\tilde{\boldsymbol{\pi}} \in \tilde{S}), \tag{B.2}$$

and $p(N \mid \boldsymbol{n}) = \int p(N \mid n, \mathcal{T}(\tilde{\boldsymbol{\pi}})) p(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) d\tilde{\boldsymbol{\pi}}$, where $p(n \mid \pi_0) = \sum_{N=n}^{\infty} L_1(N, \pi_0 \mid n) p(N)$ and $p(N \mid n, \pi_0) = L_1(N, \pi_0 \mid n) p(N) / p(n \mid \pi_0)$, with $\pi_0 = \mathcal{T}(\tilde{\boldsymbol{\pi}})$. As we discuss in Section B.2.3, we can compute $p(n \mid \pi_0)$, and thus $p(N \mid n, \pi_0)$, analytically for common priors on N. If one has access to Markov chain Monte Carlo (MCMC) samples $\{\tilde{\boldsymbol{\pi}}^{[t]}\}_{t=1}^{T}$ from $p(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$, one can then generate MCMC samples $\{N^{[t]}\}_{t=1}^{T}$ from $p(N \mid \boldsymbol{n})$ via $N^{[t]} \sim p(N \mid n, \mathcal{T}(\tilde{\boldsymbol{\pi}}^{[t]}))$. Summaries of the marginal posterior of N can then be calculated based on these samples.

B.2.2 Mixing and Matching Identifying Assumptions and Priors

While computation as described in the previous section may seem straightforward, the marginal posterior for the observed cell probabilities, $p(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$, depends on the specific combination of priors for $\tilde{\boldsymbol{\pi}}$ and N and identifying assumption $\mathcal{T}$. Thus we need new MCMC samples from $p(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$ for each new combination of priors and identifying assumption, which can be difficult both technically and computationally. Rather than develop new MCMC samplers for each combination, we will rely on a combination of existing software and a computationally cheap rejection sampler.

Let $p_C(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) \propto L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})p(\tilde{\boldsymbol{\pi}})$ denote the marginal "posterior" for the observed cell probabilities using just the conditional likelihood L_2. We use the subscript C (for "C"onditional) to denote that it is a special density that we are introducing for computational purposes. We can then rewrite the actual marginal posterior for the observed cell probabilities (B.2) as $p(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) \propto p(n \mid \mathcal{T}(\tilde{\boldsymbol{\pi}}))I(\tilde{\boldsymbol{\pi}} \in \tilde{S})p_C(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$. This suggests a computationally cheap rejection sampler to generate samples from $p(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$, if we have access to MCMC samples from $p_C(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$ (Smith and Gelfand, 1992):

1. Generate $U \sim \text{UNIF}(0, 1)$ and $\tilde{\boldsymbol{\pi}} \sim p_C(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$ independently.

2. If $U < p(n \mid \mathcal{T}(\tilde{\boldsymbol{\pi}}))I(\tilde{\boldsymbol{\pi}} \in \tilde{S})/\{\max_{\pi_0} p(n \mid \pi_0)\}$ accept $\tilde{\boldsymbol{\pi}}$. Else go back to (1).

Thus, for a given prior $p(\tilde{\boldsymbol{\pi}})$, if we want to perform prior sensitivity analyses for N and/or sensitivity analyses probing the identifying assumption as discussed in Sections 4.2 and 4.3 of the main text, we can take a one time sample from $p_C(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$, and then reuse this sample to generate samples from $p(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$ for each combination of prior for N and identifying assumption. The approach just described is only useful if we have access to MCMC samples from $p_C(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$. The rest of this section will describe how we can generate samples from the density $p_C(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$ using existing software.

Previous work in Bayesian MSE specifies priors for $\tilde{\boldsymbol{\pi}}$ indirectly. In particular, most work specifies priors on reparametrizations of the cell probabilities $\boldsymbol{\pi}$, such as log-linear models or LCMs, which induce priors for $\boldsymbol{\pi}$, and thus for $\tilde{\boldsymbol{\pi}}$. Let $p^w(\boldsymbol{\pi})$ denote what we will call the "working" prior for $\boldsymbol{\pi}$, which induces the prior $p(\tilde{\boldsymbol{\pi}})$ we would like to use. We use the superscript w (for "w"orking) to denote that it is a special density that we are introducing

for computational purposes. Consider the "working" posterior for $\boldsymbol{\pi}$, $p^w(\boldsymbol{\pi} \mid \boldsymbol{n}) \propto \sum_{N=n}^{\infty} p(\boldsymbol{n}, n_0 \mid N, \boldsymbol{\pi}) p^w(\boldsymbol{\pi})/N$, obtained using the "working" prior for N of $p^w(N) \propto 1/N$. The "posterior" for $\tilde{\boldsymbol{\pi}}$ under this working prior combination is equal to $p_C(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) \propto L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) p(\tilde{\boldsymbol{\pi}})$, as $p(n \mid \pi_0) \propto 1/n$ under the working prior for N (see Table 5). Thus, given MCMC samples, $\{\boldsymbol{\pi}^{[t]}\}_{t=1}^{T}$, drawn from $p^w(\boldsymbol{\pi} \mid \boldsymbol{n})$, letting $\tilde{\pi}_{\boldsymbol{h}}^{[t]} = \pi_{\boldsymbol{h}}^{[t]}/(1 - \pi_0^{[t]})$, $\{\tilde{\boldsymbol{\pi}}^{[t]}\}_{t=1}^{T}$ are MCMC samples drawn from $p_C(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$.

Thus if we want to use the prior $p(\tilde{\boldsymbol{\pi}})$ induced by a working prior $p^w(\boldsymbol{\pi})$, we can rely on a combination of existing software and a computationally cheap rejection sampler to generate draws from the posterior $p(N, \tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$ for any combination of prior for N and identifying assumptions, as long as the software uses the prior $p^w(N) \propto 1/N$. Note that our prior for N *does not* have to be $p^w(N)$. This is the case for most existing software, including the `R` package `conting` (Overstall and King, 2014), which implements a reversible-jump MCMC sampler to target $p^w(\boldsymbol{\pi} \mid \boldsymbol{n})$ under a working prior $p^w(\boldsymbol{\pi})$ induced by a prior that averages over all hierarchical log-linear models (King and Brooks, 2001), and the `R` package `LCMCR`, which implements a data augementation Gibbs sampler to target $p^w(\boldsymbol{\pi} \mid \boldsymbol{n})$ under a working prior $p^w(\boldsymbol{\pi})$ induced by a Dirichlet process prior for LCMs (Manrique-Vallier, 2016). The steps of the MCMC samplers used in these packages are model specific and we would not be able to use them if we tried to create bespoke MCMC samplers targeting the marginal posterior in (B.2). We note that this approach is closely related to the working prior approach of Linero (2017), with some necessary modifications specific to MSE.

B.2.3 Recommended Priors for the Population Size, N

In Table 5 we catalog $p(n \mid \pi_0) = \sum_{N=n}^{\infty} L_1(N, \pi_0 \mid n) p(N)$ and $p(N \mid n, \pi_0) = L_1(N, \pi_0 \mid n) p(N)/p(n \mid \pi_0)$ under Poisson, negative-binomial, and binomial priors for N, in addition to the class of priors $p(N) \propto (N - \ell)!/N!$, where $\ell \in \{0, 1, 2, \cdots\}$, suggested by Fienberg et al. (1999). This class of priors contains both the improper uniform prior, $p(N) \propto 1$, when $\ell = 0$, and the improper scale prior, $p(N) \propto 1/N$, when $\ell = 1$. If $p(n \mid \pi_0)$ is not available analytically, for example when $p(N)$ is beta-binomial, we recommend truncating the prior for N to the range $\{1, \cdots, N_{max}\}$ where N_{max} is an upper bound on the population size, in which case $p(n \mid \pi_0)$ can be computed numerically.

Table 5: Catalog of $p(N \mid n, \pi_0)$ and $p(n \mid \pi_0)$ under common priors for N.

Prior	$p(N)$	$p(N \mid n, \pi_0)$	$p(n \mid \pi_0)$
$\text{Pois}(M)$	$(M)^N e^{-M}/N!$	$n + \text{Pois}(\pi_0 M)$	$\text{Pois}((1-\pi_0)M)$
$\text{NB}\left(a, \frac{M}{M+a}\right)$	$\binom{N+a-1}{N}(\frac{M}{M+a})^N(\frac{a}{M+a})^a$	$n + \text{NB}\left(n+a, \frac{M\pi_0}{M+a}\right)$	$\text{NB}\left(a, \frac{(1-\pi_0)M}{(1-\pi_0)M+a}\right)$
$\text{Bin}(M, q)$	$\binom{M}{N} q^N (1-q)^{M-N}$	$n + \text{Bin}\left(M-n, \frac{\pi_0 q}{\pi_0 q + 1 - q}\right)$	$\text{Bin}(M, (1-\pi_0)q)$
Fienberg et al. (1999)	$\propto (N-\ell)!/N!$	$n + \text{NB}(n-\ell+1, \pi_0)$	$\propto \frac{(n-\ell)!}{n!}(1-\pi_0)^{\ell-1}$

The improper scale prior, under which $p(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) \propto p_C(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) I(\tilde{\boldsymbol{\pi}} \in \tilde{S})$, is a common "noninformative" prior for N and has the nice property that the posterior mean of N conditional on $\tilde{\boldsymbol{\pi}}$ is the Horvitz-Thompson estimator (Horvitz and Thompson, 1952), $n/\{1 - \mathcal{T}(\tilde{\boldsymbol{\pi}})\}$, which is well understood in the present context (see e.g. Rukhin, 1975). Recall that the Horvitz-Thompson estimator also arose when considering frequentist inference in Appendix B.1.1. Following Link (2013), we recommend using this prior in the absence of substantive knowledge about N.

When incorporating substantive knowledge about N into an informative prior for N we recommend using a negative-binomial or beta-binomial prior, as we have found Poisson and binomial priors to be more informative than we would usually like to use. For concreteness in the main text we focused on the negative-binomial prior. In Table 5, we use a common parameterization for the negative-binomial distribution in terms of the mean M and overdispersion parameter a. This parameterization arises from a Poisson-gamma mixture, where $N \mid \delta \sim \text{Poisson}(M\delta)$, $\delta \sim \text{Gamma}(a, a)$. As $a \to \infty$ the prior approaches a Poisson prior with mean M, and as $a \to 0$ the prior approaches the improper scale prior.

B.3 Regularization and Data Sparsity

As discussed in Section 3.1 of the main text, when one uses a model that places little to no restrictions on the observed data distribution (as we advocate for in Section 2.6 of the main text), this can lead to population size estimates with large variances associated with them. This typically occurs when the data is sparse, i.e. when some cells of the observed contingency table are small (or even 0). Data sparsity can be a problem when conducting frequentist analyses, as the standard asymptotic arguments used in Appendix B.1.1 to derive standard errors and

confidence intervals are generally not valid. This issue is secondary to our main focus of choosing the identifying assumption, in the sense that the amount of sparsity in the data should not affect the choice of identifying assumption.

We discuss here two possible routes to reduce the variance of population size estimators. The first route is to place restrictions on the observed data distribution, as advocated for by the quote of Fienberg (1972) in Section 3.1 of the main text. This would require the restricted model to truly hold, otherwise the lower estimated variance would not be valid and the population size estimate could be arbitrarily biased. We would generally prefer not to take this route, as such restrictions are typically hard to justify in practice (see e.g. Dellaportas and Forster, 1999; Whitehead et al., 2019). Further, even if one places correct restrictions on the observed data distribution, in a frequentist analysis the standard errors and confidence intervals derived in Appendix B.1.1 can still be invalid when the data are sparse.

The second route is to use some form of regularization when estimating the observed cell probabilities within a model that places little to no restrictions on the observed data distribution. Regularization reduces the variances of estimates, at the cost of increasing the bias of estimates, by shrinking parameter estimates to a predetermined subset of parameter space. We now briefly discuss how regularization can be incorporated into frequentist or Bayesian analyses:

- In a frequentist analysis, regularization can be incorporated through some form of penalized likelihood (Good and Gaskins, 1971), where instead of estimating the observed cell probabilities $\tilde{\boldsymbol{\pi}}$ by maximizing the conditional likelihood as described in Appendix B.1.1, one would maximize the sum of the conditional likelihood and a penalty term

$$\hat{\boldsymbol{\pi}} = \arg\max_{\tilde{\boldsymbol{\pi}} \in \tilde{S}} L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) - cJ(\tilde{\boldsymbol{\pi}}). \tag{B.3}$$

  Here J is a penalty function and $c > 0$ is a regularization parameter. When $c = 0$ the estimate corresponds to the conditional maximum likelihood estimate, and as c increases the estimate gets shrunk to some subset of $\tilde{S}$ defined by the penalty function J. It will typically be feasible to obtain estimates by solving B.3. However, deriving standard errors and confidence intervals for these estimates can be difficult, especially when the data

is sparse. Nonstandard asymptotic theory may be required (see e.g. Nardi and Rinaldo, 2012).

- In a Bayesian analysis, regularization is inherent due to the prior distribution for $\tilde{\boldsymbol{\pi}}$. Here the prior serves a similar purpose to the penalty function J in a frequentist analysis, defining the subset of $\tilde{S}$ to which estimates of $\tilde{\boldsymbol{\pi}}$ are shrunk. Note that the computational techniques in Appendix B.2 are still valid even when the data are sparse.

We note that there is a common difficulty associated with regularizing estimates in a frequentist or Bayesian analysis: choosing where to shrink estimates of the observed cell probabilities, $\tilde{\boldsymbol{\pi}}$; i.e. choosing the penalty function, J, in a frequentist analysis or the prior in a Bayesian analysis. A fruitful direction for future research is to understand what are choices of penalty functions or priors that produce population size estimates with desirable properties when the data are sparse (e.g. good frequentist performance).

C Web Appendix C: Identifying Assumption Derivations

The purpose of this appendix is to derive the identifying assumptions associated with no-highest-order interaction assumption and the K'-list marginal no-highest-order interaction assumption.

C.1 Derivation for No-Highest-Order Interaction Assumption

Recall from Section 3.1 of the main text that we have the following relationship between the cell probabilities and the highest order interaction, $\lambda_{\mathbf{1}}$: $\prod_{\boldsymbol{h}\in H}\pi_{\boldsymbol{h}}^{I_{odd}(\boldsymbol{h})}/\prod_{\boldsymbol{h}\in H}\pi_{\boldsymbol{h}}^{I_{even}(\boldsymbol{h})}=\exp\{(-1)^{K+1}\lambda_{\mathbf{1}}\}$, where $I_{odd}(\boldsymbol{h})=I(\sum_{k=1}^{K}h_k \text{ is odd})$ and $I_{even}(\boldsymbol{h})=I(\sum_{k=1}^{K}h_k \text{ is even})$. Suppose we fix $\lambda_{\mathbf{1}}\in\mathbb{R}$, or equivalently $\xi=\exp\{(-1)^{K+1}\lambda_{\mathbf{1}}\}\in\mathbb{R}^{+}$. Under this assumption we have that $\prod_{\boldsymbol{h}\in H}\pi_{\boldsymbol{h}}^{I_{odd}(\boldsymbol{h})}/\prod_{\boldsymbol{h}\in H}\pi_{\boldsymbol{h}}^{I_{even}(\boldsymbol{h})}=\xi$. Multiplying the left-hand side by $1=\left(\frac{1-\pi_0}{1-\pi_0}\right)^{2^{K-1}}$, we find that $\tilde{\Pi}_{odd}/\{[\pi_0/(1-\pi_0)]\tilde{\Pi}_{even}\}=\xi$, where $\tilde{\Pi}_{odd}=\prod_{\boldsymbol{h}\in H^*}\tilde{\pi}_{\boldsymbol{h}}^{I_{odd}(\boldsymbol{h})}$ and $\tilde{\Pi}_{even}=\prod_{\boldsymbol{h}\in H^*}\tilde{\pi}_{\boldsymbol{h}}^{I_{even}(\boldsymbol{h})}$. Rearranging terms and solving for π_0, we find that the assumption that ξ is a fixed value corresponds to the explicit functional relationship

$$\mathcal{T}(\tilde{\boldsymbol{\pi}})=\frac{\tilde{\Pi}_{odd}/\tilde{\Pi}_{even}}{\xi+\tilde{\Pi}_{odd}/\tilde{\Pi}_{even}}. \tag{C.1}$$

The identifying assumption corresponding to the no-highest-order interaction assumption is recovered by setting $\lambda_{\mathbf{1}} = 0$, or equivalently $\xi = 1$: $\mathcal{T}(\tilde{\boldsymbol{\pi}}) = (\tilde{\Pi}_{odd}/\tilde{\Pi}_{even})/(1 + \tilde{\Pi}_{odd}/\tilde{\Pi}_{even})$. The observed-data distribution is not restricted by the assumption that the highest-order interaction is fixed, and thus models that use this assumption without any extra assumptions regarding the observed cell probabilities are nonparametric identified.

C.2 Derivation for K'-list Marginal No-Highest-Order Interaction Assumption

Suppose we assume that $\prod_{\boldsymbol{g}\in G} \pi_{\boldsymbol{g}+}^{I_{odd}(\boldsymbol{g})} / \prod_{\boldsymbol{g}\in G} \pi_{\boldsymbol{g}+}^{I_{even}(\boldsymbol{g})} = \xi$, where $\xi \in \mathbb{R}^+$ is fixed. Multiplying the left-hand side by $1 = \left(\frac{1-\pi_0}{1-\pi_0}\right)^{2^{K'-1}}$, we find that $\tilde{\Pi}_{odd,+}/\{[\pi_0/(1-\pi_0) + \tilde{\pi}_{0+}]\tilde{\Pi}_{even,+}\} = \xi$, where $\tilde{\Pi}_{odd,+} = \prod_{\boldsymbol{g}\in G^*} \tilde{\pi}_{\boldsymbol{g}+}^{I_{odd}(\boldsymbol{g})}$ and $\tilde{\Pi}_{even,+} = \prod_{\boldsymbol{g}\in G^*} \tilde{\pi}_{\boldsymbol{g}+}^{I_{even}(\boldsymbol{g})}$. Rearranging terms and solving for π_0, we find that the assumption that ξ is a fixed value corresponds to the explicit functional relationship

$$\mathcal{T}(\tilde{\boldsymbol{\pi}}) = \frac{\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \xi\tilde{\pi}_{0+}}{\xi + (\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \xi\tilde{\pi}_{0+})}. \tag{C.2}$$

The identifying assumption corresponding to the K'-list marginal no-highest-order interaction assumption is recovered by setting $\xi = 1$: $\mathcal{T}(\tilde{\boldsymbol{\pi}}) = (\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \tilde{\pi}_{0+})/(1 + \tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \tilde{\pi}_{0+})$.

As noted in Section 4.3 of the main text, the the K'-list marginal no-highest-order interaction assumption does not imply that there is no highest-order interaction for all K lists, as $\prod_{\boldsymbol{h}\in H} \pi_{\boldsymbol{h}}^{I_{odd}(\boldsymbol{h})} / \prod_{\boldsymbol{h}\in H} \pi_{\boldsymbol{h}}^{I_{even}(\boldsymbol{h})} = (\tilde{\Pi}_{odd}/\tilde{\Pi}_{even}) \times (\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \tilde{\pi}_{0+})^{-1} \neq 1$ in general.

D Web Appendix D: Latent Class Model Simulations

The purpose of this appendix is conduct simulation studies demonstrating the practical implications of Theorem A.2. In particular, we present a variety of simulations exploring the frequentist properties of the Bayesian LCM of Manrique-Vallier (2016). In each example we generate 200 data sets from the model in (A.1) for a given number of lists K and a fixed parameter setting of $\theta \in \Omega_{\mathcal{Q}_J}$, i.e. a fixed population size N and a J-class LCM $Q \in \mathcal{Q}_J$. For all examples we will use $N \in \{2000, 10000, 100000\}$. For each simulated data set, we fit the Bayesian LCM of Manrique-Vallier (2016) as implemented in the `R` package `LCMCR`, using J latent classes (i.e. the same number

that generated the data) and the default prior for $\boldsymbol{\nu}$, by running the Gibbs sampler implemented in `LCMCR` for $250,000$ iterations, with the first $50,000$ tossed for burn-in. We note that `LCMCR` uses the improper scale prior for N, i.e. $p(N) \propto 1/N$, and a flat prior for $\boldsymbol{q}$, i.e. $q_{jk} \overset{i.i.d.}{\sim} \text{UNIF}(0,1)$, which can not be changed. For each parameter setting of $\theta \in \Omega_{\mathcal{Q}_J}$ we examine the frequentist performance of the posterior median, 95% credible interval, and 50% credible interval for estimating the unobserved cell probability, π_0, through the sample mean of the posterior medians, the sample coverage of the 95% credible intervals, the sample mean of the 95% credible interval widths over the 200 replications, the sample coverage of the 50% credible intervals, and the sample mean of the 50% credible interval widths over the 200 replications.

D.1 Example 1

In this example we consider data from $K = 2$ lists generated from the two-class LCM Q_{1a} with parameters given in Table 6. Under Q_{1a}, $\tilde{\pi}_{Q_{1a},(0,1)} = 0.276$, $\tilde{\pi}_{Q_{1a},(1,0)} = 0.276$, $\tilde{\pi}_{Q_{1a},(1,1)} = 0.448$, and $\pi_{Q_{1a},0} = 0.316$. There exists another two-class LCM Q_{1b}, with parameters given in Table 6, such that $\tilde{\boldsymbol{\pi}}_{Q_{1a}} = \tilde{\boldsymbol{\pi}}_{Q_{1b}}$ but $\pi_{Q_{1b},0} = 0.219$. Because $\mathcal{P}_{\Omega_{\mathcal{Q}_2}}$ is not conditionally identified when $K = 2$, if we try to perform estimation within $\mathcal{P}_{\Omega_{\mathcal{Q}_2}}$, which contains the true data generating model, there is no guarantee that we can estimate well the cell probabilities and population size which generated the data. This example was constructed using the counterexample used to prove Theorem A.2.

Table 6: Parameters of two latent class models, Q_{1a} and Q_{1b} (rounded for presentation)

	ν_1	ν_2	q_{11}	q_{12}	q_{21}	q_{22}
Q_{1a}	0.500	0.500	0.248	0.248	0.743	0.743
Q_{1b}	0.857	0.143	0.495	0.495	0.990	0.990

The results of the simulation using data generated using the LCM Q_{1a} are presented in Table 7. We see that the posterior median has a negative bias that does not vanish as N increases. One may have thought that the posterior median might possibly be a good estimator for $\pi_{Q_{1b},0} = 0.219$ since Q_{1a} and Q_{1b} induce the same observed-data distribution. However, the posterior median is also negatively biased for estimating $\pi_{Q_{1b},0}$, which suggests there

are other LCMs in $\mathcal{Q}_2$ that induce very similar observed-data distributions to Q_{1a} and Q_{1b} but with different induced unobserved cell probabilities. While the 95% credible interval has nominal coverage when $N = 2000$, as N increases, coverage decreases and is no longer nominal. The 50% credible interval have essentially 0 coverage for settings of N, even for $N = 2000$ where the 95% credible interval has nominal coverage. This suggests the 95% credible interval only has nominal coverage at $N = 2000$ due to wide tails of the posterior for N.

Table 7: Results of the simulation study where data was generated from the two-class latent class model Q_{1a}. Truth is $\pi_{Q_{1a},0} = 0.316$.

N	Mean Posterior Median	95% CI Coverage	Mean 95% CI Width	50% CI Coverage	Mean 50% CI Width
2000	0.148	0.955	0.332	0.000	0.029
10000	0.146	0.730	0.316	0.000	0.023
100000	0.151	0.265	0.167	0.055	0.037

D.2 Example 2

One may object to the practicality of Example 1, as it examined a two class LCM constructed using the counterexample from the proof of Theorem A.2, and is thus an M_h LCM. So we now consider the following example. Manrique-Vallier (2016) presented a simulation study with $K = 5$ lists where data was generated from a LCM with $J = 2$ classes, which we reproduce in Table 8. The parameters of this LCM were based on a hypothetical population where a small proportion of people have a high probability of being observed, and a large proportion of people have a small probability of being observed, which is plausible in some human rights applications.

Suppose we only observed lists three and four, so that we have data from $K = 2$ lists generated from the two-class LCM Q_2 with parameters given in Table 9. Under Q_2, $\pi_{Q_2,0} = 0.704$. Just as in the previous example, because $\mathcal{P}_{\Omega_{\mathcal{Q}_2}}$ is not conditionally identified when $K = 2$, if we try to perform estimation within $\mathcal{P}_{\Omega_{\mathcal{Q}_2}}$, which contains the true data generating model, there is no guarantee that we can estimate well the cell probabilities and population size which generated the data. The results of the simulation using data generated using the LCM Q_2 are presented in Table 10. We see that the posterior median has a large negative bias that does not vanish as N

Table 8: Parameters of latent class model which generated data in simulation of Manrique-Vallier (2016).

		Sampling probabilities, $\boldsymbol{q}$				
Class	$\boldsymbol{\nu}$	List 1	List 2	List 3	List 4	List 5
1	0.900	0.033	0.033	0.099	0.132	0.033
2	0.100	0.660	0.825	0.759	0.990	0.693

increases, while the mean 95% and 50% credible interval widths decrease as N increases. Further, the 95% and 50% credible intervals have essentially 0 coverage across all N.

Table 9: Parameters of latent class model Q_2

ν_1	ν_2	q_{11}	q_{12}	q_{21}	q_{22}
0.900	0.100	0.099	0.132	0.759	0.990

Table 10: Results of the simulation study where data was generated from the two-class latent class model Q_2. Truth is $\pi_{Q_2,0} = 0.704$.

N	Mean Posterior Median	95% CI Coverage	Mean 95% CI Width	50% CI Coverage	Mean 50% CI Width
2000	0.285	0.000	0.408	0.000	0.055
10000	0.283	0.010	0.401	0.000	0.036
100000	0.285	0.030	0.256	0.000	0.035

D.3 Example 3

In this example we present two more frequentist simulation studies based on only observing a subset of the five lists from the simulation of Manrique-Vallier (2016).

First suppose that we only observe lists two, three, and four from the simulation of Manrique-Vallier (2016), so that we have data from $K = 3$ lists generated from the two-class LCM Q_{3a} with parameters given in Table

11. Under Q_{3a}, $\pi_{Q_{3a},0} = 0.681$. Because $\mathcal{P}_{\Omega_{\mathcal{Q}_2}}$ is not conditionally identified when $K = 3$, if we try to perform estimation within $\mathcal{P}_{\Omega_{\mathcal{Q}_2}}$, which contains the true data generating model, there is no guarantee that we can estimate well the cell probabilities and population size which generated the data. The results of the simulation using data generated using the LCM Q_{3a} are presented in Table 12. We see that the posterior median has a slight negative bias that becomes negligible as N increases. The 95% credible intervals have over-coverage across the different settings of N. The 50% credible intervals have nominal coverage when $N = 2000$, but have over-coverage as N increases.

Table 11: Parameters of latent class model Q_{3a}

		Sampling probabilities, $\boldsymbol{q}$		
Class	$\boldsymbol{\nu}$	List 2	List 3	List 4
1	0.900	0.033	0.099	0.132
2	0.100	0.825	0.759	0.990

Table 12: Results of the simulation study where data was generated from the two-class latent class model Q_{3a}. Truth is $\pi_{Q_{3a},0} = 0.681$.

N	Mean Posterior Median	95% CI Coverage	Mean 95% CI Width	50% CI Coverage	Mean 50% CI Width
2000	0.622	1.000	0.339	0.510	0.120
10000	0.667	1.000	0.274	0.800	0.091
100000	0.682	1.000	0.209	0.965	0.074

Suppose now we only observe lists two, three, four, and five from the simulation of Manrique-Vallier (2016), so that we have data from $K = 4$ lists generated from the two-class LCM Q_{3b} with parameters given in Table 13. Under Q_{3b}, $\pi_{Q_{3b},0} = 0.658$. Because $\mathcal{P}_{\Omega_{\mathcal{Q}_2}}$ is conditionally identified when $K = 4$, we know that, since $\mathcal{P}_{\Omega_{\mathcal{Q}_2}}$ contains the true data generating model, we can consistently estimate the cell probabilities and population size which generated the data. The results of the simulation using data generated using the LCM Q_{3a} are presented in Table 14. We

see that the posterior median has a negative bias that becomes negligible as N increases, as expected. The 95% and 50% credible intervals have slight under-coverage when $N = 2000$, which becomes nominal as N increases.

Table 13: Parameters of latent class model Q_{3b}

		Sampling probabilities, $\boldsymbol{q}$			
Class	$\boldsymbol{\nu}$	List 2	List 3	List 4	List 5
1	0.900	0.033	0.099	0.132	0.033
2	0.100	0.825	0.759	0.990	0.693

Table 14: Results of the simulation study where data was generated from the two-class latent class model Q_{3b}. Truth is $\pi_{Q_{3b},0} = 0.658$.

N	Mean Posterior Median	95% CI Coverage	Mean 95% CI Width	50% CI Coverage	Mean 50% CI Width
2000	0.631	0.915	0.190	0.445	0.065
10000	0.653	0.940	0.089	0.505	0.031
100000	0.658	0.955	0.028	0.485	0.010

D.4 Example 4

In this example we present three more frequentist simulation studies based on adding a third class to the LCM from the simulation study of Manrique-Vallier (2016), representing a small proportion of the population having a probability of being observed somewhere between the other two classes. The parameters of this new LCM are given in Table 15.

First suppose that we only observe lists two, three, and four from the LCM in Table 15, so that we have data from $K = 3$ lists generated from the three-class LCM Q_{4a} with parameters given in Table 16. Under Q_{4a}, $\pi_{Q_{4a},0} = 0.613$. Because $\mathcal{P}_{\Omega_{\mathcal{Q}_3}}$ is not conditionally identified when $K = 3$, if we try to perform estimation within $\mathcal{P}_{\Omega_{\mathcal{Q}_3}}$, which contains the true data generating model, there is no guarantee that we can estimate well the cell

Table 15: Parameters of latent class model which generated data in simulation of Manrique-Vallier (2016), with a third class added.

		Sampling probabilities, $\boldsymbol{q}$				
Class	$\boldsymbol{\nu}$	List 1	List 2	List 3	List 4	List 5
1	0.700	0.033	0.033	0.099	0.132	0.033
2	0.200	0.275	0.250	0.200	0.300	0.325
3	0.100	0.660	0.825	0.759	0.990	0.693

probabilities and population size which generated the data. The results of the simulation using data generated using the LCM Q_{4a} are presented in Table 17. We see that the posterior median has a negative bias that does not vanish as N increases. The 95% credible intervals have over-coverage across the different settings of N, while the 50% credible intervals have under-coverage across the different settings of N. Similar to Example 1 in Section D.1, this suggests the 95% credible interval only has over-coverage due to wide tails of the posterior for N.

Table 16: Parameters of latent class model Q_{4a}

		Sampling probabilities, $\boldsymbol{q}$		
Class	$\boldsymbol{\nu}$	List 1	List 2	List 3
1	0.700	0.033	0.099	0.132
2	0.200	0.250	0.200	0.300
3	0.100	0.825	0.759	0.990

Next suppose that we only observe lists two, three, four, and five from the LCM in Table 15, so that we have data from $K = 4$ lists generated from the three-class LCM Q_{4b} with parameters given in Table 14. Under Q_{4b}, $\pi_{Q_{4b},0} = 0.569$. Because $\mathcal{P}_{\Omega_{\mathcal{Q}_3}}$ is not conditionally identified when $K = 4$, if we try to perform estimation within $\mathcal{P}_{\Omega_{\mathcal{Q}_3}}$, which contains the true data generating model, there is no guarantee that we can estimate well the cell

Table 17: Results of the simulation study where data was generated from the two-class latent class model Q_{4a}. Truth is $\pi_{Q_{4a},0} = 0.613$.

N	Mean Posterior Median	95% CI Coverage	Mean 95% CI Width	50% CI Coverage	Mean 50% CI Width
2000	0.524	1.000	0.387	0.210	0.119
10000	0.537	1.000	0.364	0.150	0.102
100000	0.538	1.000	0.323	0.175	0.096

probabilities and population size which generated the data. The results of the simulation using data generated using the LCM Q_{4b} are presented in Table 19. We see that the posterior median has a negative bias that decreases as N increases. While the 95% and 50% credible intervals do not have nominal coverage, coverage improves as N increases (but is still far from nominal even when $N = 100000$).

Table 18: Parameters of latent class model Q_{4b}

		Sampling probabilities, $\boldsymbol{q}$			
Class	$\boldsymbol{\nu}$	List 1	List 2	List 3	List 4
1	0.700	0.033	0.099	0.132	0.033
2	0.200	0.250	0.200	0.300	0.325
3	0.100	0.825	0.759	0.990	0.693

Next suppose that we observe all five lists from the LCM in Table 15, so that we have data from $K = 5$ lists generated from the three-class LCM which we will refer to as Q_{4c}. Under Q_{4c}, $\pi_{Q_{4c},0} = 0.536$. Because $\mathcal{P}_{\Omega_{\mathcal{Q}_3}}$ is not conditionally identified when $K = 5$, if we try to perform estimation within $\mathcal{P}_{\Omega_{\mathcal{Q}_3}}$, which contains the true data generating model, there is no guarantee that we can estimate well the cell probabilities and population size which generated the data. The results of the simulation using data generated using the LCM Q_{4c} are presented in Table 20. We see that the posterior median has a negative bias that decreases as N increases. While the 95% and 50% credible intervals do not have nominal coverage, coverage improves as N increases.

Table 19: Results of the simulation study where data was generated from the two-class latent class model Q_{4b}. Truth is $\pi_{Q_{4b},0} = 0.569$.

N	Mean Posterior Median	95% CI Coverage	Mean 95% CI Width	50% CI Coverage	Mean 50% CI Width
2000	0.469	0.525	0.199	0.090	0.065
10000	0.509	0.630	0.128	0.120	0.041
100000	0.519	0.695	0.066	0.290	0.023

Table 20: Results of the simulation study where data was generated from the two-class latent class model Q_{4c}. Truth is $\pi_{Q_{4c},0} = 0.536$.

N	Mean Posterior Median	95% CI Coverage	Mean 95% CI Width	50% CI Coverage	Mean 50% CI Width
2000	0.435	0.415	0.169	0.050	0.055
10000	0.500	0.875	0.132	0.370	0.044
100000	0.523	0.895	0.053	0.490	0.018

D.5 Takeaways

When using the model $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ for multiple-systems estimation, one is relying on the assumption that the data was generated from a distribution in $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$. If a practitioner is comfortable with the assumption that $2J \leq K$, then we know the model is conditionally identified, and thus this assumption is a combination of an explicit identifying assumption (which is currently unknown) and possibly some restrictions on the observed-data distribution. Due to conditional identification, the practitioners have guarantees under this assumption that they can estimate the population size, and other parameters, well if their observed sample size n is large enough. However, if a practitioner is not comfortable with this assumption, and chooses to use $J > K/2$, they have no such guarantees as they are using a model that is not conditionally identified.

Through the four example simulation studies in this appendix we saw examples where models that were not conditionally identified had good frequentist performance (Q_{3a}, Q_{4a}) and bad frequentist performance (Q_1, Q_2, Q_{4b}, Q_{4c}) according to some of our simulation summary measures. The good and bad frequentist performances

could have been due to

- where the prior of Manrique-Vallier (2016) places mass in the parameter space $\Omega_{\mathcal{Q}_J}$ (e.g. good frequentist performance if it places enough prior mass around the true data generating parameters),
- whether there actually exists other LCMs in $\mathcal{Q}_J$ that induce similar observed cell probabilities to the true data generating parameters but a different unobserved cell probability (e.g. good frequentist performance if other LCMs do not exist with these properties),
- or some combination of the two previous factors.

We currently have no way to tease apart these factors and tell when a model that is not conditionally identified will have good or bad performance. This is a problem for using these models in practice, as we have no way to tell practitioners "under these assumptions the model will perform well".

We believe there are two routes forward to combat this problem, if one wants to use LCMs for multiple-systems estimation. The first option is to further study technical results for conditional identification in LCMs. For example, as we discussed in Section A.6, suppose we can prove under further (practically relevant) restrictions on $\mathcal{Q}_J$ that $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ is conditionally identified for some $J > K/2$. We would then be able to expand the range of models we could fit under which we had guarantees that we could estimate well the parameters of the model.

The other option is to study LCMs through the framework of partial identification (Tamer, 2010; Gustafson, 2010), which was recently used in multiple-systems estimation by Sun et al. (2020) for frequentist inference for partially-identified log-linear models. This would require both: 1) a better technical understanding of what parameters, or functions of parameters, of LCMs are not identified, and 2) placing substantively meaningful priors on the non-identified parameters (i.e. priors informed by substantive knowledge concerning the population of interest and how the data was collected) if a Bayesian approach is taken. Without 1), the best we can do in a Bayesian approach is to place substantively meaningful priors on all LCM parameters, i.e. on $\Omega_{\mathcal{Q}_J}$. The prior for $\Omega_{\mathcal{Q}_J}$ of Manrique-Vallier (2016) is based on the Dirichlet Process prior specification of Dunson and Xing (2009), which is a prior of technical convenience. Specifying a substantively meaningful prior for $\Omega_{\mathcal{Q}_J}$ would require being able to

specify a prior for the class membership probabilities $\boldsymbol{\nu}$ and for the class specific observation probabilities $\boldsymbol{q}$. It is difficult to imagine a scenario in which a practicioner would have knowledge of the population of interest and how the data was collected that could be incorporated into priors for all $J(K+1)-1$ of the parameters ($\boldsymbol{\nu}$ and $\boldsymbol{q}$).

While we do not believe that latent class models cannot be used for multiple-systems estimation (see our application in Section 5 of the main text where we use the LCM prior of Manrique-Vallier (2016) to induce a prior for the observed cell probabilities $\tilde{\boldsymbol{\pi}}$), we do believe that there needs to be further research to understand under what assumptions LCMs do and do not perform well in practice. We discuss one further area of research before concluding this section. The start of this section began by assuming that a practitioner assumed their data was generated by a distribution in $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$. It is not clear to the authors how in practice one would choose a specific value of J. In practice, how would a practitioner choose between $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ and $\mathcal{P}_{\Omega_{\mathcal{Q}_{J'}}}$ for $J \neq J'$? What characteristics of the population being studied and the data collection process would allow one to differentiate between these two models? Research into understanding how to elicit plausible values of J would help to justify the use of the model $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ in practice.

E Web Appendix E: Kosovo Analysis Appendix

This appendix serves three purposes: 1) to describe the difficulty in justifying the NHOI assumption for the Kosovo data, 2) to describe a prior sensitivity analysis for the Bayesian analyses of the Kosovo data, and 3) to describe a sensitivity analysis for the Kosovo data probing the NHOI assumption.

E.1 The No-Highest-Order Interaction Assumption

The Kosovo data set has $K = 4$ lists, which we will order (without loss of generality) so that the American Bar Association Central and East European Law Initiative (ABA) list is first, the Human Rights Watch (HRW) list is second, the Organization for Security and Cooperation in Europe (OSCE) list is third, and the list constructed from exhumation reports conducted on behalf of the International Criminal Tribunal for the Former Yugoslavia (EXH) is fourth. Let $\text{Odds}(h_1 = 1 \mid h_2 = 1, h_3, h_4) = \pi_{(1,1,h_3,h_4)}/\pi_{(0,1,h_3,h_4)}$ denote the odds that an individual

is observed in list 1, conditional on being observed in list 2 and the inclusion patterns h_3, h_4 for lists 3 and 4. For example, if $h_3 = 0$ and $h_4 = 1$, $\text{Odds}(h_1 = 1 \mid h_2 = 1, h_3 = 0, h_4 = 1)$ is the odds that an individual is observed in list 1, conditional on being observed in lists 2 and 4 and not being observed in list 3. Similarly let $\text{Odds}(h_1 = 1 \mid h_2 = 0, h_3, h_4) = \pi_{(1,0,h_3,h_4)}/\pi_{(0,0,h_3,h_4)}$ denote the odds that an individual is observed in list 1, conditional on not being observed in list 2 and the inclusion patterns h_3, h_4 for lists 3 and 4. We can then define $\text{OR}(h_3, h_4) = \text{Odds}(h_1 = 1 \mid h_2 = 1, h_3, h_4)/\text{Odds}(h_1 = 1 \mid h_2 = 0, h_3, h_4)$ as the odds ratio for lists 1 and 2, conditional on the inclusion patterns h_3, h_4 for lists 3 and 4. Following Section 4.1 of the main text, the no-highest-order interaction assumption assumes that $\text{OR}(1,0)/\text{OR}(0,0) = \text{OR}(1,1)/\text{OR}(0,1)$, i.e. the highest-order interaction for the first three lists, conditional on not being observed in list 4, $\text{OR}(1,0)/\text{OR}(0,0)$, is equal to the highest-order interaction for the first three lists, conditional on being observed in list 4, $\text{OR}(1,1)/\text{OR}(0,1)$.

This assumption is obscure and hard to justify based on our knowledge of how the four lists were generated. As the validity of our analysis rests on this assumption being correct, we stress that we are not confident that this assumption holds, and thus we are not confident in the validity of the analysis of the Kosovo data set using the NHOI assumption.

E.2 Prior Sensitivity Analyses

In this section we perform prior sensitivity analyses for the Bayesian analyses of the Kosovo data from the main text. For N, we will consider the negative-binomial prior specification described in the main text, in addition to the improper scale prior discussed in Appendix B.2.3. For the observed cell probabilities $\tilde{\boldsymbol{\pi}}$, we will consider four prior specifications : 1) the prior induced from using the Dirichlet process prior of Manrique-Vallier (2016) for the J class LCM $\Omega_{LCM,J}$, with $J = 10$ and default hyperparameters, as implemented in the `R` package `LCMCR` (i.e. the prior used in the main analyses), 2) a flat Dirichlet prior, i.e. $\tilde{\boldsymbol{\pi}} \sim \text{DIRICHLET}(1, \cdots, 1)$, 3) the prior induced from using $\text{NORMAL}(0, 5^2)$ priors for the log-linear parameters in the saturated log-linear model Ω_{LL}, fit using the `Stan` probabilistic programming language (Carpenter et al., 2017), and 4) the prior induced from using the Bayesian model averaging prior of King and Brooks (2001) for the log-linear parameters in the saturated log-linear

model Ω_{LL}, with the unit information prior on log-linear parameters, as implemented in the `R` package `conting` (Overstall and King, 2014). We note that `conting` uses an alternative log-linear parameterization based on sum to zero constraints rather than corner point constraints used in Section 3.1 of the main text. For each combination of identifying assumption and priors for N and $\tilde{\boldsymbol{\pi}}$ we fit the corresponding model using the computational approach described in Appendix B.2.2.

In Table 21 we present posterior means and 95% credible intervals for N under each prior combination under the 2-list marginal NHOI assumption, i.e. assuming marginal independence of the ABA and HRW lists. The posterior density for N under each prior combination under the 2-list marginal NHOI assumption is displayed in Figure 2. For each prior for $\tilde{\boldsymbol{\pi}}$, the posterior for N does not appear to be sensitive to the prior for N, as the point estimates and credible intervals are essentially the same between the two priors for N. Across the different priors for $\tilde{\boldsymbol{\pi}}$, the posterior summaries are fairly consistent, with the posterior summaries under the LCM prior for $\tilde{\boldsymbol{\pi}}$ being slightly lower than under the other priors. We note that all of the credible intervals fall within the confidence interval of Spiegel and Salama (2000).

Table 21: Posterior means and 95% credible intervals for N under each combination of prior for N and $\tilde{\boldsymbol{\pi}}$, under the 2-list marginal NHOI assumption.

	Improper Scale Prior	Negative-Binomial
Conting	9618 [8224, 11195]	9621 [8232, 11191]
Dirichlet	9536 [8113, 11252]	9540 [8123, 11247]
LCMCR	9353 [7959, 11063]	9359 [7967, 11059]
Log-Linear	9764 [8277, 11549]	9766 [8288, 11550]

In Table 22 we present posterior means and 95% credible intervals for N under each prior combination under the NHOI assumption. The posterior density for N under each prior combination under the NHOI assumption is displayed in Figure 3. For each prior for $\tilde{\boldsymbol{\pi}}$, the posterior for N is somewhat sensitive to the prior for N, as the posterior mean and credible interval limits are always larger under the improper scale prior compared to the negative-binomial prior for N. The posterior for N appears to be the most sensitive to the prior for N under

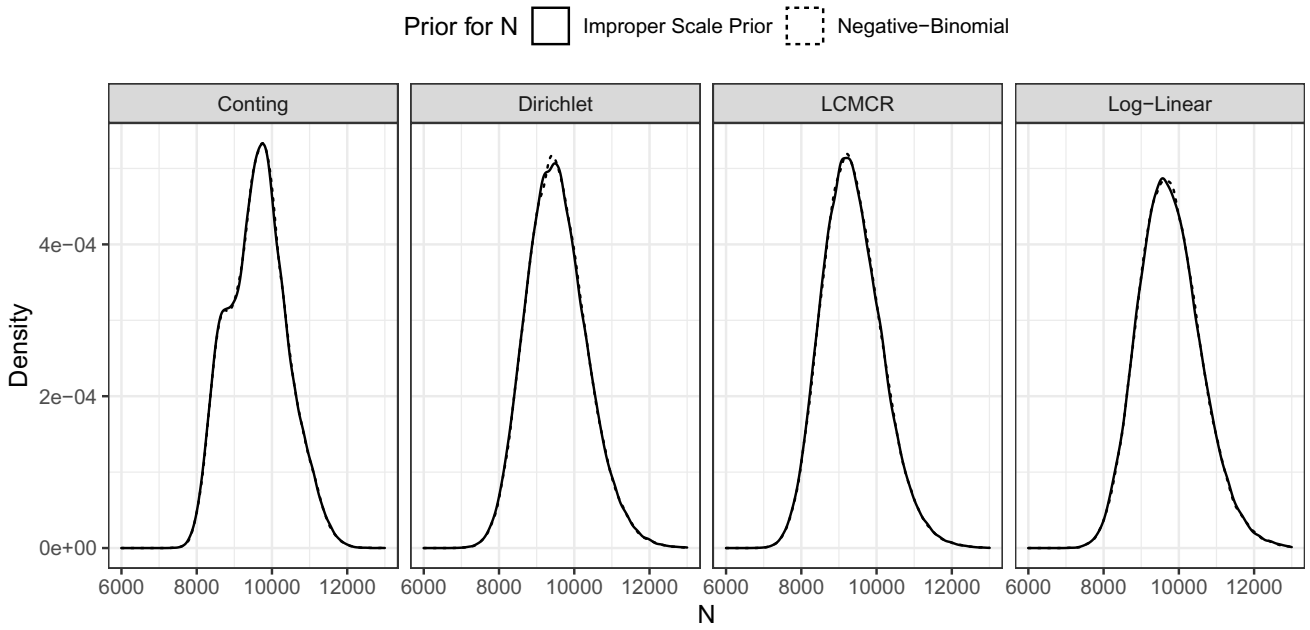

Figure 2: Posterior density of N under each combination of prior for N and $\tilde{\boldsymbol{\pi}}$, under the 2-list marginal NHOI assumption.

the Dirichlet and log-linear priors for $\tilde{\boldsymbol{\pi}}$, where the posterior means and upper credible interval limits increase by several thousand when using the improper scale prior for N instead of the negative-binomial prior. Across the different priors for $\tilde{\boldsymbol{\pi}}$, the posteriors corresponding to the Dirichlet prior, the log-linear model prior, and the LCM prior of Manrique-Vallier (2016) are in relative agreement. The posterior corresponding to the Dirichlet prior is the most diffuse of the three, and the posterior corresponding to the LCM prior of Manrique-Vallier (2016) is the most concentrated of the three. The posterior corresponding to the log-linear model prior of King and Brooks (2001), implemented in the `conting` package, is multimodal, which is not unexpected as it is performing Bayesian model averaging (Hoeting et al., 1999) over all hierarchical log-linear models. Due to this multimodality, point estimates (e.g. the posterior mean) may not be reliable summaries of the posterior distribution. We note that all of the credible intervals contain the point estimate of Spiegel and Salama (2000).

Table 22: Posterior means and 95% credible intervals for N under each combination of prior for N and $\tilde{\boldsymbol{\pi}}$, under the NHOI assumption.

	Improper Scale Prior	Negative-Binomial
Conting	13000 [9202, 19971]	12694 [9175, 19299]
Dirichlet	18500 [9402, 35908]	16051 [9098, 27679]
LCMCR	14695 [9423, 23675]	14071 [9321, 21604]
Log-Linear	16209 [8731, 30025]	14719 [8579, 24878]

E.3 A Sensitivity Analysis Probing the NHOI Assumption

We now perform a sensitivity analysis probing the no-highest-order interaction assumption. We will consider models with the identifying assumption in Section 4.2 of the main text, varying ξ over $\{1/2, 2/3, 1, 3/2, 2\}$ (following Gerritse et al., 2015). For each value of ξ, we will present both a frequentist analysis and a Bayesian analysis, with the Bayesian analysis using the same priors from the main analysis as presented in Section 5.1 of the main text. This sensitivity analysis is limited in that we followed Gerritse et al. (2015) and chose an arbitrary range of values for ξ around 1. Due to the difficulty in interpreting the highest-order interaction when there are $K = 4$ lists, we are not able to say with confidence whether this range of values is meaningful or not. In Table 23 we present the results from our frequentist and Bayesian analyses under each identifying assumption.

Table 23: Point estimates and 95% uncertainty intervals for sensitivity analysis probing the NHOI assumption. For the Bayesian analysis the point estimate is the posterior mean. In this table ξ is a ratio of ratios of odds ratios, as described in Section 4.2 of the main text and Appendix E.1.

	$\xi = 1$ / 2	$\xi = 2$ / 3	$\xi = 1$	$\xi = 3$ / 2	$\xi = 2$
Frequentist	29483 [6210, 52757]	23212 [5757, 40668]	16941 [5304, 28579]	12761 [5002, 20520]	10670 [4851, 16490]
Bayesian	21476 [13518, 33507]	17983 [11492, 27987]	14071 [9321, 21604]	11121 [7766, 16564]	9538 [6943, 13821]

The results are not very robust to misspecification of ξ in the chosen range. The uncertainty intervals when $\xi = 1/2$ and $\xi = 2$ barely overlap. For the Bayesian analysis, the posterior mean when $\xi = 2$ is 32% lower than

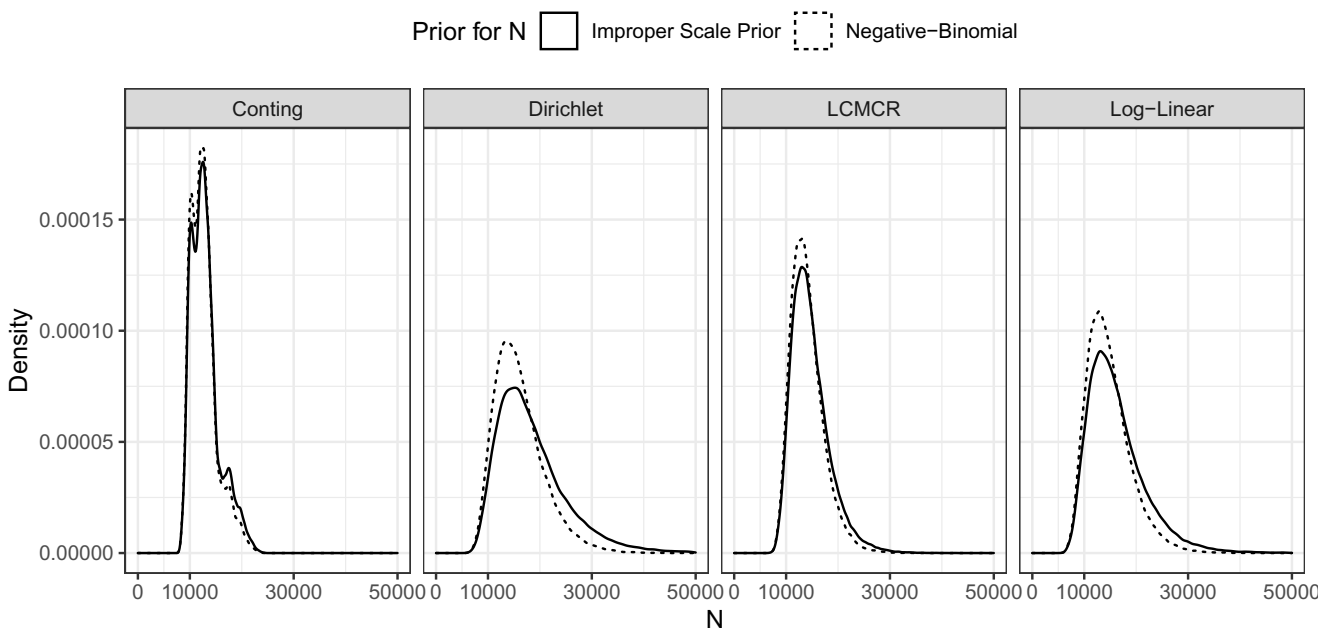

Figure 3: Posterior density of N under each combination of prior for N and $\tilde{\boldsymbol{\pi}}$, under the NHOI assumption.

the posterior mean when $\xi = 1$ (i.e. under the no-highest-order interaction assumption), the posterior mean when $\xi = 1/2$ is 53% higher than the posterior mean when $\xi = 1$, and the posterior mean $\xi = 1/2$ is more than twice the posterior mean when $\xi = 2$. These differences are even more dramatic for the frequentist analysis. This lack of robustness to misspecification of ξ would be a cause for concern if the no-highest-order interaction assumption was plausible, and the deviations from the assumption in terms of ξ were also plausible, in the context of the Kosovo data set.

References

ABA/AAAS (2000). Political killings in Kosova/Kosovo, March-June 1999. Technical report, American Bar Association Central and East European Law Initiative and the American Association for the Advancement of Science.

Agresti, A. (2003). *Categorical data analysis*, volume 482. John Wiley & Sons.

Aleshin-Guendel, S. (2020). On the Identifiability of Latent Class Models for Multiple-Systems Estimation. *arXiv preprint arXiv:2008.09865* .

Allman, E. S., Matias, C., Rhodes, J. A., et al. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics* **37,** 3099–3132.

Anderson, M. and Fienberg, S. E. (1999). *Who counts?: The politics of census-taking in contemporary America*. Russell Sage Foundation.

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

Ball, P., Betts, W., Scheuren, F., Dudukovich, J., and Asher, J. (2002). *Killings and Refugee Flow in Kosovo March-June 1999*. American Association for the Advancement of Science and American Bar Association Central and East European Law Initiative.

Bird, S. M. and King, R. (2018). Multiple systems estimation(or capture-recapture estimation) to inform public policy. *Annual review of statistics and its application* **5,** 95–118.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete multivariate analysis: theory and practice*. Springer Science & Business Media.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software* **76,**.

DasGupta, A. and Rubin, H. (2005). Estimation of binomial parameters when both n, p are unknown. *Journal of Statistical Planning and Inference* **130,** 391–404.

Dellaportas, P. and Forster, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86,** 615–633.

Dunson, D. B. and Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* **104,** 1042–1051.

Farcomeni, A. and Tardella, L. (2012). Identifiability and inferential issues in capture-recapture experiments with heterogeneous detection probabilities. *Electronic Journal of Statistics* **6,** 2602–2626.

Fienberg, S. E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika* **59,** 591–603.

Fienberg, S. E., Johnson, M. S., and Junker, B. W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **162,** 383–405.

Fienberg, S. E. and Manrique-Vallier, D. (2009). Integrated methodology for multiple systems estimation and record linkage using a missing data formulation. *AStA Advances in Statistical Analysis* **93,** 49–60.

Gelman, A., Simpson, D., and Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy* **19,** 555.

Gerritse, S. C., van der Heijden, P. G., and Bakker, B. F. (2015). Sensitivity of population size estimation for violating parametric assumptions in log-linear models. *Journal of official statistics* **31,** 357–379.

Good, I. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58,** 255–277.

Gustafson, P. (2010). Bayesian inference for partially identified models. *The International Journal of Biostatistics* **6,**.

Haberman, S. J. (1979). *Analysis of Qualitative Data. Volume 2, New Developments.* Academic Press.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science* pages 382–401.

Hogan, J. W. and Daniels, M. J. (2008). *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis.* Chapman and Hall/CRC.

Holzmann, H., Munk, A., and Zucchini, W. (2006). On identifiability in capture–recapture models. *Biometrics* **62,** 934–936.

Hook, E. B. and Regal, R. R. (1995). Capture-recapture methods in epidemiology: methods and limitations. *Epidemiologic reviews* **17,** 243–264.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association* **47,** 663–685.

Huggins, R. (2001). A note on the difficulties associated with the analysis of capture–recapture experiments with heterogeneous capture probabilities. *Statistics & probability letters* **54,** 147–152.

Iacopino, V., Frank, M. W., Bauer, H. M., Keller, A. S., Fink, S. L., Ford, D., Pallin, D. J., and Waldman, R. (2001). A population-based assessment of human rights abuses committed against ethnic Albanian refugees from Kosovo. *American Journal of Public Health* **91,** 2013–2018.

Jackson, C. (2011). Multi-state models for panel data: the msm package for R. *Journal of statistical software* **38,** 1–28.

King, R. and Brooks, S. (2001). On the Bayesian analysis of population size. *Biometrika* **88,** 317–336.

Linero, A. R. (2017). Bayesian nonparametric analysis of longitudinal studies in the presence of informative missingness. *Biometrika* **104,** 327–341.

Link, W. A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics* **59,** 1123–1130.

Link, W. A. (2006). Rejoinder to "On Identifiability in Capture-Recapture Models". *Biometrics* **62,** 936–939.

Link, W. A. (2013). A cautionary note on the discrete uniform prior for the binomial N. *Ecology* **94,** 2173–2179.

Lum, K. and Ball, P. (2015). Estimating undocumented homicides with two lists and list dependence. Technical report, Human Rights Data Analysis Group.

Madigan, D. and York, J. C. (1997). Bayesian methods for estimation of the size of a closed population. *Biometrika* **84,** 19–31.

Manrique-Vallier, D. (2016). Bayesian population size estimation using Dirichlet process mixtures. *Biometrics* **72,** 1246–1254.

Manrique-Vallier, D., Ball, P., and Sulmont, D. (2019). Estimating the Number of Fatal Victims of the Peruvian Internal Armed Conflict, 1980-2000: an application of modern multi-list Capture-Recapture techniques. *arXiv preprint arXiv:1906.04763* .

Manrique-Vallier, D., Price, M. E., and Gohdes, A. (2013). Multiple systems estimation techniques for estimating casualties in armed conflicts. *Counting civilian casualties: An introduction to recording and estimating nonmilitary deaths in conflict* pages 165–182.

Nardi, Y. and Rinaldo, A. (2012). The log-linear group-lasso estimator and its asymptotic properties. *Bernoulli* **18,** 945–974.

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife monographs* pages 3–135.

Overstall, A. and King, R. (2014). conting: An R package for Bayesian analysis of complete and incomplete contingency tables. *Journal of Statistical Software* **58,** 1–27.

Pezzott, G. L. M., Salasar, L. E. B., Leite, J. G., and Louzada-Neto, F. (2019). A note on identifiability and maximum likelihood estimation for a heterogeneous capture-recapture model. *Communications in Statistics-Theory and Methods* pages 1–21.

Regal, R. R. and Hook, E. B. (1991). The effects of model selection on confidence intervals for the size of a closed population. *Statistics in Medicine* **10,** 717–721.

Regal, R. R. and Hook, E. B. (1998). Marginal versus conditional versus 'structural source' models: a rationale for an alternative to log-linear methods for capture-recapture estimates. *Statistics in medicine* **17,** 69–74.

Rukhin, A. (1975). Statistical decision about the total number of observable objects. *Sankhyā: The Indian Journal of Statistics, Series A* pages 514–522.

Sadinle, M. (2018). Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations. *The Annals of Applied Statistics* **12,** 1013–1038.

Sanathanan, L. (1972). Estimating the size of a multinomial population. *The Annals of Mathematical Statistics* pages 142–152.

Silverman, B. (2020). Multiple systems analysis for the quantification of modern slavery: Classical and Bayesian approaches. *Journal of the Royal Statistical Society, Series A* **183,** 691–736.

Smith, A. F. and Gelfand, A. E. (1992). Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician* **46,** 84–88.

Spiegel, P. B. and Salama, P. (2000). War and mortality in Kosovo, 1998–99: an epidemiological testimony. *The Lancet* **355,** 2204–2209.

Sun, J., Van Baelen, L., Plettinckx, E., and Crawford, F. W. (2020). Partial identification and dependence-robust confidence intervals for capture-recapture surveys. *arXiv preprint arXiv:2008.00127* .

Tahmasebi, B., Motahari, S. A., and Maddah-Ali, M. A. (2018). On the Identifiability of Finite Mixtures of Finite Product Measures. *arXiv preprint arXiv:1807.05444* .

Tamer, E. (2010). Partial identification in econometrics. *Annu. Rev. Econ.* **2,** 167–195.

Whitehead, J., Jackson, J., Balch, A., and Francis, B. (2019). On the Unreliability of Multiple Systems Estimation for Estimating the Number of Potential Victims of Modern Slavery in the UK. *Journal of Human Trafficking* pages 1–13.