Bayesian Population Size Estimation Using Dirichlet Process Mixtures

Daniel Manrique-Vallier*

January 20, 2016

Abstract

We introduce a new Bayesian non-parametric method for estimating the size of a closed population from multiple-recapture data. Our method, based on Dirichlet process mixtures, can accommodate complex patterns of heterogeneity of capture, and can transparently modulate its complexity without a separate model selection step. Additionally, it can handle the massively sparse contingency tables generated by large numbers of recaptures with moderate sample sizes. We develop an efficient and scalable MCMC algorithm for estimation. We apply our method to simulated data, and to two examples from the literature of estimation of casualties in armed conflicts.

Keywords: Capture-Recapture; Casualties in conflicts; Dirichlet process mixtures; Latent class models; Model selection.

1 Introduction

Starting with the work of Sanathanan (1972) and Fienberg (1972), and the development of methods for analyzing discrete multivariate data, the literature in multiple-recapture

*Daniel Manrique-Vallier is Assistant Professor at the Department of Statistics, Indiana University, Bloomington, IN 47408 (e-mail: dmanriqu@indiana.edu).

and multiple systems estimation has devoted considerable attention to departures from the basic assumptions of independence and homogeneity prevalent in earlier work; e.g. Darroch (1958); Cormack (1968). In particular, several methods designed to accommodate individual heterogeneity structures have been proposed (e.g. Sanathanan, 1972; Fienberg et al., 1999; Manrique-Vallier and Fienberg, 2008). Similar proposals have also been developed in the animal estimation literature (e.g. Norris and Pollock, 1996).

Finite mixture models are a special class of methods for heterogeneous populations that arises naturally as the representation of the aggregation of two or more distinct homogeneous subpopulations. Their application to multiple-recapture (e.g. Norris and Pollock, 1996; Basu and Ebrahimi, 2001) can be considered model-based analogous to stratified multiple-recapture estimation (e.g. Sekar and Deming, 1949) in the absence of covariate information. These models can in principle accommodate arbitrarily complex joint distributions, as long as the number of mixture components is adequately selected (Vermunt et al., 2008).

In this paper we introduce a fully Bayesian multiple-recapture method based on Dirichlet process mixtures of product-Bernoulli distributions, originally introduced by Dunson and Xing (2009) as a general-purpose method for modeling sparse contingency tables, and modified for handling structural zeros by Manrique-Vallier and Reiter (2014). This model has some similarities with finite mixture models but differs in that it does not require the specification of the number of mixture components in advance. Instead, it considers an infinite number of them, and favors a data-learned sparse representation by placing most of the probability mass into a small finite subset. It also bears some similarities with approaches based on model averaging (e.g. Madigan and York, 1997; Arnold et al., 2010) in that it incorporates dimensionality uncertainty into estimates. A major advantage of our method is its computational expediency, and its robustness in conditions of extreme cell sparsity.

The rest of this article is organized as follows. In Section 2 we describe the general problem of closed population multinomial multiple-recapture estimation from a missing data perspective. In Section 3 we introduce the Bayesian Non-Parametric Latent Class model (NPLCM) as a flexible alternative for modeling heterogeneity, and adapt it to the multiple-recapture problem. In Section 4 we outline an efficient Markov chain Monte Carlo sampling algorithm for posterior simulation from our model. In Section 5 we apply our method to simulated data and to two examples taken from the literature on estimation of casualties in armed conflicts. We conclude with a discussion about the limitations of our method, some potential uses, and possible extensions.

2 General Framework for Multinomial Multiple-Recapture Estimation

Building on ideas from Fienberg and Manrique-Vallier (2009), we frame multiple-recapture estimation as a missing data problem (Little and Rubin, 2002). Let us consider a closed finite population of N individuals. We assume that each individual can be either listed or missed by any of J lists that partially enumerate that population. We write $x_{ij} = 1$ to indicate that individual $i \in \{1, ..., N\}$ was captured by list $j \in \{1, ..., J\}$, and $x_{ij} = 0$ to indicate otherwise. We group these *capture indicators* into individual *capture vectors*, $\mathbf{x}_i = (x_{i1}, ..., x_{iJ}) \in \{0, 1\}^J$. For example, a capture vector $\mathbf{x}_i = (0, 1, 0, 0)$ denotes an individual that has only been recorded in list $j = 2$, from $J = 4$ lists. We observe that any individual with a capture vector composed uniquely of zeros, $\mathbf{0} = (0, ..., 0)$, is by definition unobserved, and therefore cannot be present in any (combined or otherwise) sample. Let $n = \sum_{i=1}^{N} I(\mathbf{x}_i \neq \mathbf{0})$ be the number of observed individuals. Here $I(\cdot)$ takes the value 1 if the condition in the argument is true and 0 otherwise. Our task is to determine the number of *unobserved* individuals, $n_0 = \sum_{i=1}^{N} I(\mathbf{x}_i = \mathbf{0})$ or, equivalently, the population size $N = n + n_0$.

Following missing data ideas, we consider a complete data generation process and a nonignorable missing data mechanism. Let the complete data generation process be $f(\mathbf{x}|\theta)$ for $\mathbf{x} \in \{0,1\}^J$, such that $\mathbf{x}_i \overset{iid}{\sim} f(\mathbf{x}|\theta)$ for $i = 1, ..., N$ with N known. The corresponding missing data mechanism consists of simply not observing individuals with a capture vector $\mathbf{0}$. Reordering the sequence of $\mathbf{x}_i$s so that all the unobservable capture vectors are grouped together at the end of the sequence—this is, at positions $i = n+1, ..., N$—we get

$$p(\mathcal{X} \mid \theta, N) = \binom{N}{n} f(\mathbf{0}|\theta)^{N-n} \prod_{i=1}^{n} f(\mathbf{x}_i|\theta) I(N \geq n), \tag{1}$$

where $\mathcal{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)$. Here we use the symbol $p(\cdot)$ to denote the density or probability mass function of the argument, to be deduced from context. In the multiple-recapture problem both N and θ are unknown and have to be estimated. We do this by specifying a prior distribution $p(N, \theta)$ and computing $p(N, \theta|\mathcal{X}) \propto p(N, \theta)p(\mathcal{X}|\theta, N)$.

3 Modeling Heterogeneity Using the Latent Class Model

The complete-data generating distribution, $f(\mathbf{x}|\theta)$, summarizes our assumptions about the characteristics of the sampling process. For example, the independence model,

$$f(\mathbf{x}|\lambda_1, ..., \lambda_J) = \prod_{j=1}^{J} \lambda_j^{x_j} (1 - \lambda_j)^{1-x_j}, \tag{2}$$

is often an adequate representation of the capture vector distribution when it is known that listing processes are independent. When this assumption holds reasonably well, methods based on (2) usually produce good estimates of the population size. However, when this is not the case, these models are notorious for producing unreliable estimates and misstating the sampling variability (Fienberg, 1972; International Working Group for Disease Monitoring and Forecasting, 1995; Fienberg et al., 1999).

An often recommended strategy when independence assumptions do not hold appropriately is to take advantage of some form of stratification scheme (Sekar and Deming, 1949; Fienberg, 1972). The idea is to segment the population into relatively homogeneous classes where simple models, like independence, can be expected to hold better; then apply those simple models on each individual stratum to produce estimates of the population size. Of course this approach is only possible when the covariate information needed to construct such stratification exists, and is closely related to the source of sampling heterogeneity.

In the absence of an appropriate stratification scheme we can consider it missing data. Let us assume that the population admits a partition into K homogeneous strata, such that the independence model holds within each of them. Let $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)$ with $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k > 0$ be the vector of strata probabilities. Then the probability mass function of the capture vectors is

$$p(\mathbf{x}|\boldsymbol{\lambda}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \lambda_{jk}^{x_j} (1 - \lambda_{jk})^{1-x_j}, \tag{3}$$

where $\boldsymbol{\lambda} = (\lambda_{jk})$ with $\lambda_{jk} \in (0, 1)$. This is a mixture model for which each component is an independence model like (2), with stratum-specific parameters. As with any other mixture, it admits an augmented data representation as the two-step process:

$$\begin{aligned} x_j|z &\overset{indep}{\sim} \text{Bernoulli}(\lambda_{jz}) \quad \text{for } j = 1, ..., J \\ z &\sim \text{Discrete}\left(\{1, 2, ..., K\}, (\pi_1, ..., \pi_K)\right). \end{aligned} \tag{4}$$

Here z is a latent variable that explicitly represents stratum assignment.

The mixture of product-Bernoulli distributions in (3) is known as the Latent Class Model (LCM; Goodman, 1974; Haberman, 1979). LCMs are often used as model-based clustering devices for discovering and characterizing latent sub-populations within a heterogeneous population, based on multivariate discrete observable attributes. Similar to other latent variable strategies, LCM modeling assumes that dependency between coordinates of the observed response vector, $\mathbf{x}$, can be fully explained by the introduction of an unobserved variable—in this case $z \in \{1, ..., K\}$ (Haberman, 1979).

Perhaps more importantly for our purposes, LCMs are also useful as general-purpose models for contingency tables with arbitrarily complex patterns of dependence between variables, even if they are not directly motivated as the representation of a stratified data generation process (e.g. Vermunt et al., 2008; Si and Reiter, 2013). It is known (Dunson and Xing, 2009) that the mixture in (3) can represent any possible discrete distribution in the space $\{0, 1\}^J$. For it to be useful in applications, however, we still need to solve the model selection problem of choosing an appropriate number of latent classes, K.

3.1 The Non-parametric LCM

Dunson and Xing (2009) proposed a Bayesian nonparametric extension to the LCM that overcomes the need of specifying the number of mixture components in advance, while simultaneously enforcing data-learned sparsity into the mixture. Instead of trying to find a "best" finite number of latent classes, they proposed to use an infinite number of them simultaneously, combined with a prior specification that induces sparsity into the mixture by concentrating most of the probability mass into a small finite subset. The resulting model, an infinite-dimensional mixture of product-multinomial distributions, retains the simplicity and expressiveness of the original LCM, but avoids the model selection problem of having to find an appropriate number latent classes, K. Additionally, it acts as a model averaging device that propagates the uncertainty of the model dimensionality into estimates.

The non-parametric LCM (henceforth NPLCM) from Dunson and Xing (2009) is a Dirichlet process mixture of product-Bernoulli distributions. It can be described through the hierarchical generative process

$$\begin{aligned} x_j|z &\overset{indep}{\sim} \text{Bernoulli}(\lambda_{jz}) \quad \text{for } j=1,...,J \\ z &\sim \text{Discrete}\left(\{1,2,...\},(\pi_1,\pi_2,...)\right) \\ \lambda_{jk} &\overset{iid}{\sim} \text{Beta}(1,1) \quad \text{for } j=1,...,J \text{ and } k=1,2,... \\ (\pi_1,\pi_2,...) &\sim \text{SB}(\alpha) \\ \alpha &\sim \text{Gamma}(a,b), \end{aligned} \tag{5}$$

where SB(α) is the stick-breaking process with parameter $\alpha > 0$ (Sethuraman, 1994). The stick breaking specification for $\boldsymbol{\pi}$ has the effect of concentrating the bulk of the probability mass into its first few coordinates, thus inducing sparsity into the mixture and avoiding overfitting. The prior distribution on α allows us to estimate the effective dimensionality of the mixture from the data; see Gelman et al. (2013) p. 553, for a discussion of this specification.

In practice, we use a finite-dimensional approximation whereby we select a large-enough upper bound for the number of latent classes, K^* (Ishwaran and James, 2001). We define the finite-dimensional stick-breaking prior, $(\pi_1,...,\pi_{K^*}) \sim \text{SB}_{K^*}(\alpha)$, by making $\pi_k = V_k \prod_{h<k}(1-V_h)$ for $V_{K^*} = 1$ and $V_1,...,V_{K^*-1} \overset{iid}{\sim} \text{Beta}(1,\alpha)$. We note that this truncation is only a computationally convenient approximation; we do not consider K^* to be a parameter in the model as in a regular finite mixtures. As we will see in the examples, as long as most of the posterior probability mass ends up concentrated on a subset of components smaller than K^*, the exact value of K^* does not have any noticeable impact on the estimates.

As Si and Reiter (2013) demonstrate in an application involving contingency tables with more than 10^{30} cells, this specification can model complex and very high-dimensional discrete joint distributions in a parsimonious way. It also has the computational advantage of allowing the use of the blocked Gibbs algorithm from Ishwaran and James (2001), and the extension for truncated supports from Manrique-Vallier and Reiter (2014).

3.2 The NPLCM Multiple-Recapture Model

Returning to our original problem of estimating the unknown size of a closed population, we obtain a joint model for the observable sample (this is, those units with capture patterns different from $\mathbf{0}$) by plugging the LCM probability mass function from equation (3) into the general multiple-recapture multinomial model described in (1),

$$p(\mathcal{X}|\boldsymbol{\lambda},\boldsymbol{\pi},N) \propto \binom{N}{n}\left[\sum_{k=1}^{K^*}\pi_k\prod_{j=1}^{J}(1-\lambda_{jk})\right]^{N-n}\prod_{i=1}^{n}\sum_{k=1}^{K^*}\pi_k\prod_{j=1}^{J}\lambda_{jk}^{x_{ij}}(1-\lambda_{jk})^{1-x_{ij}}. \qquad (6)$$

Using the latent variable representation from (4), we see that (6) is equivalent to a marginalized version of the augmented data representation

$$p(\mathcal{X},\mathbf{z},\mathbf{z}^0|\lambda,\pi,N) \propto \binom{N}{n}\prod_{i=1}^{n_0}\pi_{z_i^0}\prod_{j=1}^{J}(1-\lambda_{jz_i^0})\times\prod_{i=1}^{n}\pi_{z_i}\prod_{j=1}^{J}\lambda_{jz_i}^{x_{ij}}(1-\lambda_{jz_i})^{1-x_{ij}}, \qquad (7)$$

where $\mathbf{z}=(z_1,...,z_n)$ and $\mathbf{z}^0=(z_1^0,...,z_{n_0}^0)$, and both z_i and z_i^0 take values on the set $\{1,...,K^*\}$ for each $i=1,\dots,n$.

We complete a full Bayesian specification by choosing prior distributions for parameters $\boldsymbol{\pi}$, and N. Following advice from Dunson and Xing (2009) and Si and Reiter (2013) we choose $\alpha\sim\text{Gamma}(0.25,0.25)$ as a default diffuse specification. We comment more on this choice in Section 5. For the total population size, we choose the improper discrete prior distribution $p(N)\propto 1/N$. This specification, which corresponds to the Jeffreys' prior for N, has several advantages. First, it has been shown to produce good results, and to avoid some paradoxical behaviors of the posterior (Link, 2013). Second, it makes the posterior distribution estimates match (after renormalization) those of an NPLCM truncated at the $\mathbf{0}$ cell (Manrique-Vallier and Reiter, 2014). Finally, it results in simple closed-form Gibbs steps in the algorithm we propose in the next section. As usual, other specifications are possible, if prior knowledge is available. In particular, as Fienberg et al. (1999) note, prior distributions of the form $p(N) \propto (N-l)!(N!)^{-1}$ for $l = 1, ..., N-1$ comprise important special cases and lead to simple conditional posterior distributions.

4 Estimation Via Markov Chain Monte Carlo

The augmented data representation in (7) leads naturally to MCMC algorithms based Gibbs sampling schemes that exploit the conditional independence given latent variables structure. However, the fact that the length of the vector $\mathbf{z}^0$ is exactly equal to $n_0 = N - n$, entails additional difficulties. Since N is itself a parameter to estimate, it is not possible to construct valid Gibbs sampling algorithms by simply deriving full conditional distributions for N and each z_i^0, because it would result in a reducible Markov chain. This difficulty was identified by Basu and Ebrahimi (2001), who proposed to overcome it by jointly sampling N and the latent variables (in this case $\mathbf{z}^0$) using a conditional decomposition. This idea was also exploited by Fienberg et al. (1999) and Manrique-Vallier and Fienberg (2008) in the context of multiple-recapture, and adapted and extended by Manrique-Vallier and Reiter (2014) as general method to sample from the NPLCM subject to complex structural zero restrictions.

In principle, the general technique from Manrique-Vallier and Reiter (2014) is directly applicable to this problem. However, the special structure of the multiple-recapture problem, where only one cell is unobservable, allows for further simplifications. Let $\boldsymbol{\omega} = (\omega_1, ..., \omega_{K^*})$, with $\omega_k = \sum_{i=1}^{n_0} I(z_i^0 = k)$. Here ω_k denotes the number of unobserved individuals that belong to latent class k. Then we get the representation

$$p(\mathcal{X}, \mathbf{z}, \boldsymbol{\omega}|\boldsymbol{\lambda}, \boldsymbol{\pi}, N) = \binom{N}{n, \omega_1, \cdots, \omega_{K^*}} \prod_{k=1}^{K^*} \left(\pi_k \prod_{j=1}^{J} (1 - \lambda_{jk}) \right)^{\omega_k} \\ \times \prod_{i=1}^{n} \pi_{z_i} \prod_{j=1}^{J} \lambda_{jz_i}^{x_{ij}} (1 - \lambda_{jz_i})^{1-x_{ij}} \times I\left(\sum_{k=1}^{K^*} \omega_k = N - n \right). \tag{8}$$

It is easy to see that (8) is equivalent to (6) after marginalizing over $\mathbf{z}$ and $\boldsymbol{\omega}$.

Now we show how to construct a Gibbs sampler algorithm for obtaining samples from the posterior distribution of parameters for this model, including the population size, N. Here we use the prior distributions proposed in Section 3.2. The first steps are similar to those proposed by Manrique-Vallier and Reiter (2014).

1. **Sample from** $p(\mathbf{z}|...)$: For $i = 1, ..., n$, sample $z_i \sim \text{Discrete}(\{1, ..., K^*\}, (p_1, ..., p_{K^*}))$, with $p_k \propto \pi_k \prod_{j=1}^{J} \lambda_{jk}^{x_{ij}} (1 - \lambda_{jk})^{1-x_{ij}}$.

2. **Sample from** $p(\boldsymbol{\lambda}|...)$: For $j = 1, \ldots, J$ and $k = 1, \ldots, K^*$ let $n_k = \sum_{i=1}^{n} I(z_i = k)$ and $n_{jk} = \sum_{i=1}^{n} I(x_{ij} = 1, z_i = k)$. Then sample $\lambda_{jk} \sim \text{Beta}\left(n_{jk} + 1, n_k - n_{jk} + \omega_k + 1\right)$.

3. **Sample from** $p(\boldsymbol{\pi}|...)$: For $k = 1, ..., K^* - 1$ sample

$$V_k \sim \text{Beta}\left(1 + \nu_k, \alpha + \sum_{h=k+1}^{K^*} \nu_h\right)$$

   where $\nu_k = n_k + \omega_k$. Let $V_K^* = 1$ and make $\pi_k = V_k \prod_{h<k}(1 - V_h)$ for all $k = 1, ..., K^*$.

4. **Sample from** $p(\alpha|...)$: $\alpha \sim \text{Gamma}(a - 1 + K^*, b - \log \pi_{K^*})$

5. **Sample from** $p(N, \boldsymbol{\omega}|...)$: The full joint conditional distribution of $\boldsymbol{\omega}$ and N is

$$p(\boldsymbol{\omega}, N|\boldsymbol{\lambda}, \mathbf{z}, \alpha, \boldsymbol{\pi}, \mathcal{X}) \propto p(N) \frac{n_0!}{\omega_1! \cdots \omega_{K^*}!} \rho_1^{\omega_1} \cdots \rho_{K^*}^{\omega_{K^*}} \times I(N = n + n_0) \tag{9}$$

where $n_0 = \sum_{k=1}^{K^*} \omega_k$ and $\rho_k = \pi_k \prod_{j=1}^{J}(1-\lambda_{jk})$. For $P(N) \propto 1/N$ this is a negative multinomial distribution—note we are not conditioning on N, and that N is completely determined by $\boldsymbol{\omega}$. Thus we obtain samples from this distribution by compounding a negative binomial with a multinomial distribution (Sibuya et al., 1964):

(a) Sample $n_0 \sim \text{NegBinomial}\left(n, 1-\sum_{k=1}^{K^*} \pi_k \prod_{j=1}^{J}(1-\lambda_{jk})\right)$. Make $N = n + n_0$.

(b) Sample $(\omega_1, ..., \omega_{K^*}) \sim \text{Multinomial}(n_0, (p_1, ..., p_{K^*}))$ for $p_k \propto \rho_k$.

We note that in spite of the complexity of the NPLCM, the proposed Gibbs sampler algorithm is remarkably simple, consisting only of sampling steps from standard distributions. Additionally, it only requires sampling $K^* \times (J+2) + n + 1$ variates per iteration. This makes it highly scalable, both on the number of lists, J, and sample size. We also note that, different from other Bayesian latent variable proposals (Fienberg et al., 1999; Basu and Ebrahimi, 2001; Manrique-Vallier and Fienberg, 2008) whose computational cost depends on the unknown n_0, the computational cost of this algorithm is fixed.

5 Example Applications

We applied our method to simulated data and to two multiple-recapture datasets taken from the literature on casualty estimation in armed conflicts. The first real dataset consists of four incomplete overlapping lists documenting killings during the Kosovo war (Ball et al., 2002). The second is a collection of 15 matched lists documenting killings due to political violence in Casanare, Colombia (Lum et al., 2010). In the Supplemental materials online we include an additional example using a classical small-sample dataset from the ecological literature.

For comparison, we also fit the independence and, where possible, more complex log-linear models (Fienberg, 1972). We do not expect simple independence models to provide good estimates due to the strong capture heterogeneity in our examples. Rather, we take them as an indication of its severity. The other log-linear models have been selected by minimizing the BIC index over the whole class of hierarchical log-linear models. This replicates a common practice in applied multi-list problems in human populations (see e.g. Hook and Regal, 2000), and in some sense seeks to give the class log-linear models the best possible chance of fitting the data and producing reasonable estimates. We performed the model search using the routines included in the package Rcapture (Baillargeon and Rivest, 2007).

In all examples we used the prior distributions described in Section 3.2. We have performed additional experiments investigating the sensitivity of our method to different choices of hyper-parameters in the prior specification α. ‘Our results, which can be consulted in the Supplemental Materials online, show that posterior inference about N is relatively insensitive to the broad range of prior specifications on α that we examine. We have set the maximum number of latent classes, K^*, in each example so that a fair number of them were estimated with negligible probability. A practical way for choosing such K^* is to initially set it to be the number of unique capture histories, and then decrease it, if possible, ensuring that still a fair number of the estimated latent classes still end up estimated with negligible probability. We note that under such conditions increasing K^* results largely in the same estimates of N. For summarizing the posterior distribution of the NPLCM we have used posterior medians as point estimates, and equal-tail 95% posterior intervals as interval estimators.

5.1 Simulated Data

We simulated the repeated multiple-recapture of a heterogeneous population of $N = 2000$ units by $J = 5$ lists. To induce heterogeneity we randomly split the population into two strata, with proportions 0.9 and 0.1, and allowed each list to include each individual with different probabilities, depending on stratum membership. We set stratum-level capture probabilities so that units in the largest stratum have small chance (between 0.033 and 0.132)

Table 1: Listing parameters for simulation experiment. For $N = 2000$ the expected observed sample size is $E[n] = 727$.

		List capture probabilities				
Stratum	Proportion	List 1	List 2	List 3	List 4	List 5
1	0.9	0.033	0.033	0.099	0.132	0.033
2	0.1	0.660	0.825	0.759	0.990	0.693

to get included, while individuals on the other have much larger probabilities (between 0.66 and 0.99); see Table 1. This specification mimics a commonly found situation in human populations where most people have a relatively small probability of being listed, but there exists a small subset (famous people, for instance) for which this probability is much higher. We repeated this procedure 200 times to obtain 200 different multiple-recapture samples. Note that according to this design the expected number of observed individuals in each sample is $E[n] = 727$, which represents only about 36.4% of the actual population size.

We have also fit independence log-linear models to the stratified data, one per stratum, using the actual assignment labels. We expect this procedure to produce the best possible estimates of the total population size, as it basically reproduces the data-generation process, and directly removes the source of the heterogeneity. It also benefits from information unavailable to the other models. For this reason the quality of these estimates is in principle unattainable by the other models. We have computed these estimates as means of obtaining an upper bound on the quality of the estimates that we can expect to get from these data.

Table 2 summarizes our results. All quantities are averages over 200 trials. As expected, the independence model fits the data poorly, severely underestimating the population size with an average estimate of 825.1. It also produces deceptively tight 95% confidence intervals, which result in an empirical coverage rate of 0%. BIC-selected log-linear models also perform poorly, this time overestimating the true counts by a factor of more than six and producing 95% confidence intervals that also fail to cover the true value every time. Log-linear models also performed poorly in terms of their mean squared error, which in both cases were several

Table 2: Summary of simulation results over 200 experiments with $N = 2000$ and $E[n] = 727$. (*)Note that 'Stratified" estimates have been obtained using stratification information unavailable to the other models.

Model	Mean $\hat{N}$	Mean CI width	MSE	Empirical Coverage
NPLCM	1935.8	868.83	49038.44	0.92
Independence	825.1	55.032	1381033.6	0
BIC-log linear	13432.0	988.49	164355253.2	0
Stratified (*)	2014.8	796.126	43099.8	0.96

orders of magnitude larger than that of the stratified model.

The NPLCM model, in contrast, produced very good estimates with a mean estimate of 1935.2. Credible intervals are short, and have an empirical coverage rate of 92%. These results are, as expected, inferior to those obtained through stratified estimation—marked with an asterisk in Table 2. However they are very close. In fact, the mean squared error of the NPLCM estimates is only about 14% larger than that from the stratified estimates—cf. 3104% and 381236% for the independence and BIC-selected models, respectively.

The remarkable performance of the NPLCM in this example might come as no surprise to some, given that we have generated the data through a stratified procedure akin to an LCM. However we note that the actual data generation process involved exactly two strata, while the NPLCM was fitted with a much higher upper truncation level ($K^* = 10$ in the full experiment; $K^* = 30$ in some initial tests). This means that in order to produce these results, the NPLCM has extracted underlying dimensionality of the mixture directly from the data, without any extra information.

5.2 Killings in Kosovo

Ball et al. (2002) used multiple-recapture data to estimate the number of casualties during the Kosovo war from March 20 to June 22, 1999. We reproduce this dataset in Table 3. It consists of $J = 4$ lists that jointly document $n = 4400$ observed killings. Using a parsimonious

Table 3: Kosovo data ($J = 4, n = 4400$). Source: Ball et al. (2002).

				List 1			
				Yes		No	
				List 2		List 2	
				Yes	No	Yes	No
List 3	Yes	List 4	Yes	27	32	42	123
			No	18	31	106	306
	No	List 4	Yes	181	217	228	936
			No	177	845	1131	??

log-linear model, Ball et al. (2002) estimated that a total of $\hat{N} = 10356$ (95% confidence interval $[9002, 12122]$) people were killed during that period.

This dataset makes an interesting test case because of the existence of an additional near-census enumeration. As part of an ambitious documentation project named the "Kosovo Memory Book", the Serbia-based Humanitarian Law Center (HLC) has compiled what some believe (Krüger and Ball, 2014; Spagat, 2014) to be a near-exhaustive list of casualties. We have used these data (Humanitarian Law Center, 2014) to produce a list of casualties for the same period considered by Ball et al. (2002). This results in a total of $N_{HLC} = 10401$ casualties. We note that this is a simple count, not a multiple-recapture estimate. We consider this number to be our benchmark for evaluation.

Table 4 details our results. The independence model again performed poorly, underestimating the number of casualties as $\hat{N} = 7393$, with a 95% interval that does not cover the benchmark. The BIC-based model selection procedure fared better, producing the estimate $\hat{N} = 10335$, similar to the count from the HLC and to the estimate from Ball et al. (2002).

The NPLCM produced excellent results, with a point estimate $\hat{N} = 10442$. Figure 1 shows the posterior distribution of the population size N. We see that the posterior dispersion of N is rather large, with a 95% credible interval $[9020, 13637]$. This interval is wider than that of the BIC selected log-linear model, at [8994, 12110]. Even though these two intervals have been constructed under different considerations, it is interesting to note that

Table 4: Estimates from examples with real data. Row marked (*) correspond to the census-like count from the Humanitarian Law Center.

Example	Model	$\hat{N}$	$CI_{95\%}$
Kosovo	HLC count (*)	10401	—
($n = 4400, J = 4$)	NPLCM	10442	[9020.0, 13637.0]
	Indendence	7392.9	[7147.7, 7656.2]
	BIC-Loglin ([123][14][34])	10334.5	[8994.3, 12110.4]
Casanare	NPLCM	4734	[4073, 5887]
($n = 2629, J = 15$)	Indendence	3733	[3605, 3873]

this discrepancy should be expected. Confidence intervals for multiple-recapture estimates are usually constructed conditioning on a selected model. This means that the BIC log-linear estimates ignore the sampling variability inherent to the data-based model selection. In contrast, the NPLCM model simultaneously estimates all the parameters of the model, including the population size and the effective dimensionality of the mixture.

5.3 Killings in Casanare

Lum et al. (2010) analyzed multi-list data documenting $n = 2629$ deaths due to political violence in the Casanare department, Colombia. This dataset has also being analyzed by Mitchell et al. (2013), and a previous version by Guzmán et al. (2007). We reproduce the data in the Supplemental Materials online. Different from most cases found in the casualty estimation literature, this dataset has been constructed from a large number of independently collected lists, $J = 15$. Using a Bayesian model averaging approach, considering only three lists at a time, Lum et al. (2010) estimated $\hat{N} = 5832$ (95% interval $[3822, 9332]$) killings.

The large number of lists in this example presents unique challenges. In principle, a large number of lists should benefit the estimation efforts, as it can potentially provide great detail on the patterns of recapture heterogeneity. However, a large number of lists also imply that, for typical sample sizes, many of the possible capture patterns will not be

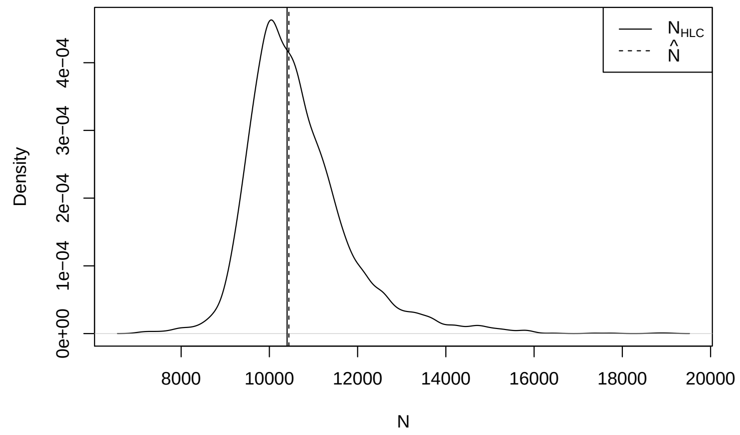


Figure 1: Posterior distribution of total population size for Kosovo data. Discontinuous vertical line show the point estimate (posterior median) $\hat{N} = 10442$; Continuous vertical line shows the near-census enumeration form the Humanitarian Law Center at $N_{HLC} = 10401$

observed. In this example $J = 15$ lists define a contingency table with $2^{15} = 32768$ cells, from which only 70 are actually present in the sample (see Supplemental Materials). This situation of extreme sparsity, in which 99.79% of the contingency table is empty, makes the use of traditional approaches, for example those based on Maximum likelihood estimation like log-linear models, extremely difficult. Additionally, the large dimensionality increases the number of possible models to fit to the data, making the model selection issue an even more pressing problem: not only the pool of possible models (for example, the family of hierarchical log-linear models) is impractically large to assess exhaustively, but the usual tests of fit are no longer valid because of the breakdown of the usual asymptotic approximations based on the χ^2 distribution due to sparsity (Fienberg and Rinaldo, 2007).

Table 4 summarizes our results. The independence model leads to the estimate $\hat{N} = 3733$, with a 95% confidence interval [4073, 5887]. Surprisingly, this simple model appears to fit the data well, with a deviance of only 1555 on 32751 degrees of freedom. However, repeating the assessment, this time using Pearson's X^2 as our fit statistic—which also has a limiting χ^2_{32751}

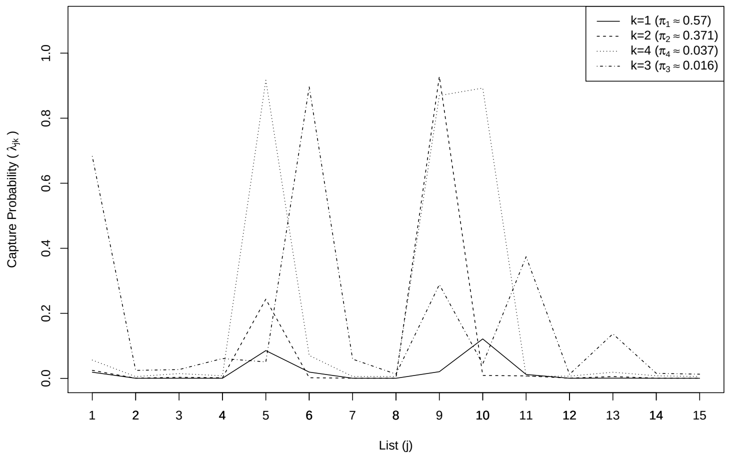
distribution—we reach the opposite conclusion, with $X^2 \approx 2.146 \times 10^{11}$ on the same degrees of freedom. This apparent contradiction points to the breakdown of asymptotic approximations based on the χ^2 distribution in sparse high-dimensional contingency tables, and thus to the difficulties involved in assessing model fit using traditional techniques. In a similar way, we were unable to perform the BIC-based model selection because sparsity prevented us from fitting most complex log-linear models. These difficulties have also plagued previous analyses, which had to be performed either selecting subsets lists at a time or collapsing some of them (Guzmán et al., 2007; Lum et al., 2010).

In contrast, we fitted the NPLCM using all the $J = 15$ lists simultaneously without problems. This led to the estimate $\hat{N} = 4734$, with a 95% credible interval [4073, 5887]. Figure 2 shows the first four mixture components identified by our model, which account for an estimated 99.46% of the total strata probability mass. From this plot it is clear that the NPLCM has identified a highly heterogeneous structure within the data, where probabilities of capture by lists vary greatly from one stratum to the other.

6 Discussion

As Dunson and Xing (2009) show, the Bayesian NPLCM has full support on the $(2^J - 1)$-dimensional probability simplex, and is therefore consistent for estimating probabilities in any contingency table. Thus, having shown the effectiveness and computational convenience of our method, we may be tempted to see the it as a sort of silver bullet that allows assumption-free multiple-recapture estimation. Unfortunately such belief would be unfounded. In general, "non-parametric multiple-recapture estimation" is somewhat of a misnomer. As Fienberg (1972) warns, multiple-recapture estimation—as any other extrapolation technique—relies on the untestable assumption that the model that describes the observed counts also applies to the unobserved ones. This makes the problem essentially

Figure 2: Posterior estimates of mixture-component capture probabilities for Casanare data. Only showing the first four components with the highest posterior probabilities (99.46% of total strata probability mass).



non-identifiable in the sense that we can produce infinite models that assign different probabilities for the unobserved cell $\mathbf{0}$ and the same set of probabilities to cells in $\{0,1\}^J \setminus \{\mathbf{0}\}$. We note that this is an intrinsic limitation of multiple-recapture estimation, regardless of the particular implementation. This was also observed by Link (2003) in the simpler context of binomial estimation with heterogeneity.

This intrinsic limitation notwithstanding, we believe that our method should be valuable tool for practitioners working in a broad class of problems. In particular, the interpretation of the NPLCM as the aggregation of homogeneous strata matches our intuition about the data generation process in many applied situations, especially in human populations. For example, in our examples estimating casualties in conflicts it is reasonable to attribute most of the dependence between lists to individual characteristics—like social visibility, or party allegiance—which make specific individuals more or less prone to be listed. We can think of similar situations in related domains like epidemiology and census corrections.

Although all of our examples have been applications to human populations, this method should in principle also work in animal abundance estimation problems. As an example, we have developed an illustrative application with a small sample size of $n = 68$ which we detail in the Supplemental materials section. However, we should note some particularities of the animal case that could make our approach not an optimal one. Most capture-recapture datasets in animal abundance estimation are the product of carefully designed experiments, in which investigators have complete control over several of the variables that drive the sampling. This allows them make some reasonably strong assumptions about the structure of the joint distribution of capture patterns. In contrast, the NPLCM is a distributionally agnostic method, which tries to estimate the joint distribution directly from the data as much as it is possible. This makes it a very attractive alternative for the estimation of elusive human populations using matched found data. However, in ecology problems—which typically deal with small sample sizes and large numbers of recaptures—we might be wasting inferential power in trying to fit such a general model, when we could get more traction by assuming a stronger structure, such as the sub-models of the M_{thb} family (Otis et al., 1978).

Our method relies only on capture pattern information for both characterizing heterogeneity, and estimating the total population size. A natural extension would be to adapt it to incorporate covariate information when it is available. This could be done by making the latent-class weights directly dependent on the covariates using, for example, covariate-dependent stick-breaking formulations (e.g. Dunson and Park, 2008); or by introducing such dependency through the latent class listing probabilities (the λs). We note that, in either case, such an extension would have account for the fact that covariates for non observed individuals are also unknown (see e.g. Tounkara and Rivest, 2015). This will be the focus of future research.

Acknowledgments

The author thanks Patrick Ball, Jule Krüger and Tamy Guberek from the Human Rights Data Analysis Group for sharing the Casanare dataset, and for help with the Kosovo data. Also to Shira Mitchell for valuable suggestions.

References

Arnold, R., Hayakawa, Y., and Yip, P. (2010), "Capture-Recapture estimation usign finite mixtures of arbitrary dimension," *Biometrics*, 66, 644–655.

Baillargeon, S. and Rivest, L.-P. (2007), "Rcapture: Loglinear Models for Capture-Recapture in R," *Journal of Statistical Software*, 19.

Ball, P., Betts, W., Scheuren, F., Dudukovic, J., and Asher, J. (2002), "Killings and Refugee Flow in Kosovo, MarchJune, 1999," Report to ICTY.

Basu, S. and Ebrahimi, N. (2001), "Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence," *Biometrika*, 88, 269–279.

Cormack, R. (1968), "The statistics of capture-recapture methods," *Oceanographic and Marine Biology Annual Review*, 6, 455–501.

Darroch, J. (1958), "The multiple-recapture census: I. Estimation of a closed population," *Biometrika*, 343–359.

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

Dunson, D. B. and Park, J.-H. (2008), "Kernel stick-breaking processes," *Biometrika*, 95, 307–323.

Fienberg, S. (1972), "The Multiple recapture census for closed populations and incomplete

2^k contingency tables," *Biometrika*, 59, 591–603.

Fienberg, S., Johnson, M., and Junker, B. (1999), "Classical multilevel and Bayesian approaches to population size estimation using multiple lists," *Journal of the Royal Statistical Society. Series A*, 162, 383–406.

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

Fienberg, S. E. and Rinaldo, A. (2007), "Three centuries of categorical data analysis: Loglinear models and maximum likelihood estimation," *Journal of Statistical Planning and Inference*, 137, 3430–3445.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian data analysis*, CRC press, 3rd ed.

Goodman, L. A. (1974), "Exploratory latent structure analysis using both identifiable and unidentifiable models," *Biometrika*, 61, 215–231.

Guzmán, D., Guberek, T., Hoover, A., and Ball, P. (2007), "Missing People in Casanare," Benetech. `https://hrdag.org/wp-content/uploads/2013/02/casanare-missing-report.pdf`, [retrieved Dec. 1st, 2014].

Haberman, S. J. (1979), *Analysis of qualitative data. vol. 2, new developments*, Chicago and London: The University of Chicago Press.

Hook, E. B. and Regal, R. R. (2000), "Accuracy of alternative approaches to capture-recapture estimates of disease frequency: internal validity analysis of data from five sources," *American Journal of Epidemiology*, 152, 771–779.

Humanitarian Law Center (2014), "The Kosovo Memory Book Project 1998–2000: List of killed, missing and disappeared 1998-2000." `http://www.kosovskaknjigapamcenja.org/db/kkp_en/index.html` [retrieved Nov 22, 2014].

International Working Group for Disease Monitoring and Forecasting (1995), “Capture-recapture and multiple-record systems estimation II: Applications in human diseases,” *American Journal of Epidemiology*, 142, 1059–1068.

Ishwaran, H. and James, L. F. (2001), “Gibbs sampling for stick-breaking priors,” *Journal of the American Statistical Association*, 96, 161–173.

Krüger, J. and Ball, P. (2014), “Evaluation of the Database of the Kosovo Memory Book,” Human Rights Data Analysis Group. `https://hrdag.org/wp-content/uploads/2013/01/Evaluation_of_the_Database_KMB-2014.pdf` [Retrieved March 20, 3015].

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

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

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data: Second Edition*, New York: John Wiley & Sons.

Lum, K., Price, M., Guberek, T., and Ball, P. (2010), “Measuring Elusive Populations with Bayesian Model Averaging for Multiple Systems Estimation: A Case Study on Lethal Violations in Casanare, 1998-2007,” *Statistics, Politics and Policy*, 1.

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. and Fienberg, S. (2008), “Population size estimation using individual level mixture models,” *Biometrical Journal*, 50, 1051–1063.

Manrique-Vallier, D. and Reiter, J. P. (2014), “Bayesian Estimation of Discrete Multivariate Latent Structure Models with Structural Zeros,” *Journal of Computational and Graphical Statistics*, 23, 1061–1079.

Mitchell, S., Ozonoff, A., Zaslavsky, A. M., Hedt-Gauthier, B., Lum, K., and Coull, B. A.

(2013), “A Comparison of Marginal and Conditional Models for CaptureRecapture Data with Application to Human Rights Violations Data,” *Biometrics*, 1022–1032.

Norris, J. and Pollock, K. (1996), “Nonparametric MLE under two closed capture-recapture models with heterogeneity,” *Biometrics*, 52, 639–649.

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*, 3–135.

Sanathanan, L. (1972), “Models and estimation methods in visual scanning experiments,” *Technometrics*, 14, 813–829.

Sekar, C. C. and Deming, W. E. (1949), “On a Method of Estimating Birth and Death Rates and the Extent of Registration,” *Journal of the American Statistical Association*, 44, 101–115.

Sethuraman, J. (1994), “A constructive definition of Dirichlet priors,” *Statistica Sinica*, 4, 639–650.

Si, Y. and Reiter, J. P. (2013), “Nonparametric Bayesian Multiple Imputation for Incomplete Categorical Variables in Large-Scale Assessment Surveys,” *Journal of Educational and Behavioral Statistics*, 38, 499–521.

Sibuya, M., Yoshimura, I., and Shimizu, R. (1964), “Negative multinomial distribution,” *Annals of the Institute of Statistical Mathematics*, 16, 409–426.

Spagat, M. (2014), “A Triumph of Remembering: Kosovo Memory Book,” `http://www.kosovomemorybook.org/wp-content/uploads/2015/02/Michael-_Spagat_Evaluation_of_the_Database_KMB_December_10_2014.pdf` [Retrieved 03/20/2015].

Tounkara, F. and Rivest, L.-P. (2015), “Mixture regression models for closed population capturerecapture data,” *Biometrics*, 71, 721–730.

Vermunt, J. K., Ginkel, J. R. V., der Ark, L. A. V., and Sijtsma, K. (2008), “Multiple imputation of incomplete categorical data using latent class analysis," *Sociological Methodology*, 38, 369–397.