Nonparametric Density Estimation for Capture-Recapture

Zachary Kurtz
Department of Statistics
Carnegie Mellon University
zkurtz@stat.cmu.edu
Version: September 14, 2012

Abstract

Capture-recapture (CRC) is a way to estimate the size of a population by combining multiple incomplete lists of population units. Accurate estimators must model dependence between lists. One kind of dependence is *unit-level list dependence*, in which previous capture directly reduces the probability of subsequent capture. Another kind of dependence arises indirectly from the heterogeneity in capture probabilities across units. Existing nonparametric CRC methods do not allow both kinds of dependence to depend on covariates. We fill this gap with a new two-stage approach. In the first stage, we estimate the conditional densities of the capture pattern as a function of the covariates. In the second stage, we impute the conditional density of the unobserved capture pattern (no captures) by applying a log-linear models locally. A Horvitz-Thompson style population estimator follows.

1 Introduction

The capture-recapture (CRC) problem is to estimate the size of some population from multiple lists. The CRC literature is large, with at least five review articles in the 1990's alone. Populations studied using CRC are diverse, including various animal species, human populations, the set of people with a disease, and the set of computer coding errors in a body of code. Despite the size of the field, nonparametrics has seen little application in the CRC research community. We propose a novel method that clears the way for nonparametrics to dominate the field of CRC.

To explain how nonparametrics can be applied in CRC, we begin with a brief introduction to the CRC problem. Every CRC estimator takes as input k different incomplete lists $L_1, ..., L_k$ of the population units. In the simplest CRC setting, there are two lists. Assume that units are perfectly matched across lists, so we have the cross-classification of units according to list membership as displayed in Table 1. Here c_{ij} is shorthand for $c_{(i,j)}$, the count of units with capture pattern (i, j).

Table 1:

		List 2	
		yes	no
List 1	yes	c_{11}	c_{10}
	no	c_{01}	c_{00}

For example, c_{10} is the number of units on List 1 but not on List 2. The number of units that are not observed on either list, c_{00}, is not observable, and estimating the population size amounts to estimating c_{00}. With three lists, the task is to estimate c_{000}, and so on.

The Petersen estimator is $\hat{c}_{00} = \frac{c_{10}c_{01}}{c_{11}}$, which can be formalized as a maximum-likelihood estimator under certain assumptions [1]. Perhaps the strongest of these assumptions is that the lists are independent; the event that a unit is captured on the first list is independent of the event that a unit is captured on the second list. However, two kinds of dependence between lists are common. The first kind of dependence is *unit-level list dependence*, in which previous capture directly reduces the probability of subsequent capture. The second kind of dependence arises indirectly as a consequence of *heterogeneity*, or variability in capture probabilities across units [2].

Both sources of dependence may depend on covariates such as Age, and much of the CRC literature in the last three decades partially addresses this fact. Logistic regression models for heterogeneity were developed for the two-list scenario [3], and extended to k lists with a simple respondent fatigue effect using conditional likelihood estimation [4]. Nonparametric regression approaches exist for the two-list scenario [5]. A grade-of-membership model for heterogeneity has been proposed, but a potential shortcoming is that the model assumes unit-level independence [6].

The method we propose applies nonparametrics to allow both heterogeneity-induced and unit-level list dependence to depend on covariates in a smooth way. We begin by estimating the conditional probability of each capture pattern as a function of the covariates; this stage can be completely nonparametric. Next, we impute the conditional density of the unobserved capture pattern (no captures) by applying log-linear models locally. The imputed conditional density implies an estimate of the detection probability for each unit; the final step is to aggregate the unit-level estimates using a Horvitz-Thompson estimator. The following sections present our approach in greater detail. Throughout, we assume that the population is closed, precluding the possibility of births, deaths, and migration.

2 Methods

Let $i = 1, ..., n_c$ index the units that are on at least one list. Let $\mathcal{N} = \{1, ..., n\}$ index the population, so that $n_c \leq n$. For each $i \in \mathcal{N}$, let $m_i := I(i \in \cup_j L_j)$ so that $n_c = \sum_{i=1}^{n} m_i$.

For each unit i and list L_j, let $y_{ij} = I(i \in L_j)$. Then $y_{i\cdot} = (y_{i1}, ..., y_{ik})$, and $y_{\cdot\cdot}$ is the $n \times k$ matrix with ith row $y_{i\cdot}$. The vector $y_{i\cdot}$ is called the *capture pattern* of the ith unit. Let $x_{i\cdot}$ denote a $1 \times q$ vector of covariates associated with the ith unit, and $x_{\cdot\cdot}$ is the $n \times q$ matrix with ith row $x_{i\cdot}$. For each $i > n_c$, the pair $(x_{i\cdot}, y_{i\cdot})$ is not observed. If $x^c_{\cdot\cdot}$ is the matrix formed by the first n_c rows of $x_{\cdot\cdot}$, and $y^c_{\cdot\cdot}$ is the matrix formed by the first n_c rows of $y_{\cdot\cdot}$, then the [observable] data consists of the pair of matrices $(x^c_{\cdot\cdot}, y^c_{\cdot\cdot})$. We will refer to the pair $(x_{\cdot\cdot}, y_{\cdot\cdot})$ as the *extended* data.

Let $\mathcal{Y}_k$ denote the set of binary row vectors of length k, so each $y_{i\cdot}$ is an element of $\mathcal{Y}_k$. Let $c_{\mathbf{y}} := |\{i : y_{i\cdot} = \mathbf{y}\}|$. The array $\mathbf{c} := \{c_{\mathbf{y}}\}_{\mathbf{y} \in \mathcal{Y}_k}$ is the contingency table of counts of units in the lists. In particular, $c_{\mathbf{0}} = n - n_c$, the number of units that are not observed on any list. Assume that $y_{i\cdot}$ is a realization of a random vector $Y_{i\cdot}$. Then, the matrix $y_{\cdot\cdot}$ is a realization of a random matrix $Y_{\cdot\cdot}$. Similarly, $\mathbf{c}$ and m_i are realizations of random quantities $\mathbf{C}$ and M_i. Let $p(i, \mathbf{y}) = P(Y_{i\cdot} = \mathbf{y})$, the probability that unit i has capture pattern $\mathbf{y}$. Then $p(i, y_{i\cdot}) = P(Y_{i\cdot} = y_{i\cdot})$.

Assume that a smooth function $r(\mathbf{y}, \mathbf{x})$ exists with the property that $p(i, y_{i\cdot}) = r(y_{i\cdot}, x_{i\cdot})$ holds for all $i \in \mathcal{N}$. Define the detection function $\psi(\mathbf{x}) = 1 - r(\mathbf{0}, \mathbf{x})$, which can be interpreted as the probability that a unit with covariates $\mathbf{x}$ will appear in at least one of the lists. The Horvitz-Thompson (HT) estimator of the population size n takes the form

$$\tilde{n} = \sum_{i:M_i=1} \frac{M_i}{\psi(x_{i\cdot})} = \sum_{i:M_i=1} \frac{1}{\psi(x_{i\cdot})} \tag{1}$$

The HT estimator relies on the detection probabilities for only the units that are observed. To use the HT estimator, we must estimate the detection function ψ. If ψ is known, the HT estimator has some nice asymptotic properties. It is easy to verify that $E\tilde{n} = n$. Moreover, $\tilde{n}$ is consistent and asymptotically normal if $\psi(x_{i\cdot})$ is uniformly bounded away from 0 and 1 for all $i \in \mathcal{N}$ [3].

The HT estimator has been applied for the CRC problem using a variety of estimators for the detection function ψ. We will propose a particularly general framework that subsumes most of the existing estimators. To this end, we define function $\pi(\mathbf{y}, \mathbf{x}) := \frac{r(\mathbf{y},\mathbf{x})}{\sum_{\mathbf{z} \neq \mathbf{0}} r(\mathbf{z},\mathbf{x})} = \frac{r(\mathbf{y},\mathbf{x})}{\psi(\mathbf{x})}$. The interpretation of π depends on its arguments. For each $\mathbf{y} \in \mathcal{Y}_k \setminus \{\mathbf{0}\}$, one can think of $\pi(\mathbf{y}, \mathbf{x})$ as the

conditional probability that a unit with covariates $\mathbf{x}$ will have capture pattern $\mathbf{y}$ *given* that the unit does not have capture pattern $\mathbf{0}$.

On the other hand, $\pi(\mathbf{0}, \mathbf{x})$ does not have the same probabilistic interpretation. Instead, $\pi(\mathbf{0}, \mathbf{x})$ is simply $r(\mathbf{0}, \mathbf{x})$ scaled up so that π is proportional to r for each fixed $\mathbf{x}$. Hence $r(\mathbf{y}, \mathbf{x}) = \psi(\mathbf{x})\pi(\mathbf{y}, \mathbf{x})$, and

$$\psi(\mathbf{x}) = \frac{\sum_{\mathbf{y}\neq\mathbf{0}} r(\mathbf{y}, \mathbf{x})}{\sum_{\mathbf{y}} r(\mathbf{y}, \mathbf{x})} = \frac{\sum_{\mathbf{y}\neq\mathbf{0}} \psi(\mathbf{x})\pi(\mathbf{y}, \mathbf{x})}{\sum_{\mathbf{y}} \psi(\mathbf{x})\pi(\mathbf{y}, \mathbf{x})} = \frac{\sum_{\mathbf{y}\neq\mathbf{0}} \pi(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{0}, \mathbf{x}) + \sum_{\mathbf{y}\neq\mathbf{0}} \pi(\mathbf{y}, \mathbf{x})} \tag{2}$$

For each fixed capture pattern $\mathbf{y}$, we think of $\pi(\mathbf{y}, \mathbf{x})$ simply as a function of $\mathbf{x}$. For all observable capture patterns, we gather these functions into an array, $\Pi^* := \{\pi(\mathbf{y}, \mathbf{x})\}_{\mathbf{y}\neq 0}$. The functions in Π^* are interesting because they are estimable directly from the data using any kind of binary regression.

Equation 2 makes it seem natural to break the process of estimating ψ into two stages. The first stage is to generate the estimates $\hat{\Pi}^* := \{\hat{\pi}(\mathbf{y}, \mathbf{x})\}_{\mathbf{y}\neq 0}$, while the second stage is to impute $\pi(\mathbf{0}, \mathbf{x})$ from $\hat{\Pi}^*$. We deal with each of these stages separately in the following sections.

2.1 Estimating Π^*

Suppose that each vector $x_{i\cdot}$ is a realization of some random variable $\mathbf{X}$. Suppose $f_M(\mathbf{x})$ is a function that satisfies $f_M(x_{i\cdot}) = P(\mathbf{X} = x_{i\cdot}|M_i = 1)$, and let $g_M(\mathbf{y}, \mathbf{x}) := \pi(\mathbf{y}, \mathbf{x})f_M(\mathbf{x})$. Then, for all $i \in \mathcal{N}$,

$$g_M(y_{i\cdot}, x_{i\cdot}) = P(Y_{i\cdot} = y_{i\cdot}|X = x_{i\cdot}, M_i = 1)P(X = x_{i\cdot}|M_i = 1) = P(y_{i\cdot}, x_{i\cdot}|M_i = 1).$$

Note that g_M and f_M each can be estimated directly from the observable data (i.e., units with $M_i = 1$), and from the definition of g_M we can express the conditional density of capture pattern $\mathbf{y}$ given $\mathbf{X} = \mathbf{x}$ as

$$\pi(\mathbf{y}, \mathbf{x}) \quad = \quad \frac{g_M(\mathbf{y}, \mathbf{x})}{f_M(\mathbf{x})}.$$

A nonparametric conditional density estimator that selects bandwidths by cross-validation has been proposed [7]. As an alternative to conditional density estimation, one could fit a separate binary regression to estimate each density in Π^*.

2.2 Imputing $\pi(\mathbf{0}, \mathbf{x})$

Log-linear models provide a flexible framework for modeling complex interactions between lists. The application of log-linear models in CRC typically begins with a *homogeneity* assumption, which says that $p(i_1, \mathbf{y}) = p(i_2, \mathbf{y}) =: p(\mathbf{y})$ for all $i_1, i_2 \in \mathcal{N}$. Next, independence between units is assumed, such that the array of counts $\mathbf{c}$ is a realization of a multinomial random variable $\mathbf{C}$ with probability array $\{p(\mathbf{y})\}_{\mathbf{y}\in\mathcal{Y}_k}$. For some parametric function f with $\sum_{\mathbf{y}} \exp f(\mathbf{y}; \theta) = 1$, one can write the capture probabilities in terms of the capture pattern: $\log p(\mathbf{y}) = f(\mathbf{y}; \theta)$.

However, given that the $c_{\mathbf{0}}$ cell of $\mathbf{c}$ is not observed, the *saturated model* has only $2^k - 2$ free parameters. Restricting attention to hierarchical models, the highest-order interaction between lists is not identifiable. Therefore, the saturated hierarchical log-linear model for three lists may be written in the form

$$\log P(Y_{i\cdot} = \mathbf{y}) \quad = \quad u + u_1\mathbf{y}_1 + u_2\mathbf{y}_2 + u_3\mathbf{y}_3 + u_{12}\mathbf{y}_1\mathbf{y}_2 + u_{13}\mathbf{y}_1\mathbf{y}_3 + u_{23}\mathbf{y}_2\mathbf{y}_3. \tag{3}$$

As given in [8], a maximum likelihood solution for the saturated model is

$$\hat{c}_{000} = e^{\hat{u}} \quad = \quad \frac{c_{111}c_{001}c_{010}c_{100}}{c_{011}c_{110}c_{101}}. \tag{4}$$

Now equation 3 can be rewritten in *local* form as an [optionally] separate model for each unit:

$$\log \pi(\mathbf{y}, x_{i\cdot}) \quad = \quad u(x_{i\cdot}) + u_1(x_{i\cdot})\mathbf{y}_1 + u_2(x_{i\cdot})\mathbf{y}_2 + \cdots + u_{23}(x_{i\cdot})\mathbf{y}_2\mathbf{y}_3. \tag{5}$$

By analogy to Equation 4, we propose the following *local log-linear* imputation equation:

$$\hat{\pi}((0,0,0), \mathbf{x}) \quad = \quad \frac{\hat{\pi}((1,1,1), \mathbf{x})\hat{\pi}((0,0,1), \mathbf{x})\hat{\pi}((0,1,0), \mathbf{x})\hat{\pi}((1,0,0), \mathbf{x})}{\hat{\pi}((0,1,1), \mathbf{x})\hat{\pi}((1,1,0), \mathbf{x})\hat{\pi}((1,0,1), \mathbf{x})}. \tag{6}$$

3 Results

A simulation experiment illustrates the method. A population of 2000 units (say, people) has Age assigned roughly consistent with the Age distribution of people. Bernoulli draws are used to generate three lists of units according the capture probabilities defined as ad-hoc continuous functions of Age, with a mean capture probability on the order of 0.6. For individuals captured on the first list, the probability of subsequent capture is reduced by about 0.1; similarly, individuals captured on the second list have a reduced probability of appearing on the third list.

For each of 25 replications of the simulation, we used the `np` package in `R` (see [7]) to estimate the conditional densities Π^* . The estimated densities for a single replication are displayed as stacked black curves in the first panel of Figure 1.

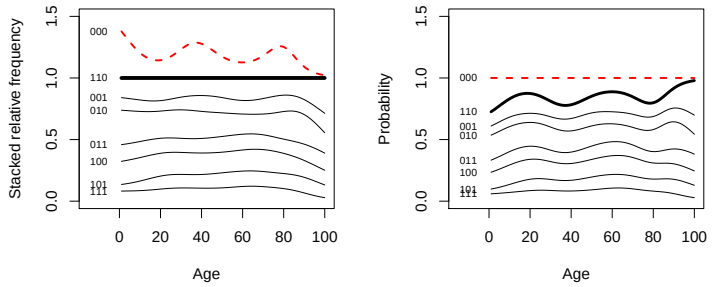

Figure 1: Stacked conditional densities for a single replication. For example, the curve labeled "100" represents the sum $\hat{\pi}((1,0,0),Age) + \hat{\pi}((1,0,1),Age) + \hat{\pi}((1,1,1),Age)$. The dashed curve represents the imputed value of $\hat{\pi}((0,0,0),Age)$. The curves are normalized in the right panel; the bold solid curve represents the detection function.

For each $i = 1, ..., n_c$, we use Equation 6 to impute $\pi(\mathbf{0}, \mathbf{x})$. This imputation is plotted as the red dashed curve in Figure 1, stacked on top of the other conditional densities. The bold curve in the right panel of Figure 1 represents the detection function $\hat{\psi}(Age)$ implied by Equation 2, and an HT estimate is immediate.

The result is biased; in 25 replications, population estimates ranged from 2005 to 2204 with a median of 2101. A primary reason for the bias is that the saturated hierarchical log-linear model is not adequate. In particular, the model does not include the highest-order interaction, which is significant due to the form of the respondent fatigue effect in the third list, which applies for all units that appear in at least one of the first two lists. The estimation problem becomes much simpler if the simulation is modified so that the fatigue effect for List 3 depends only on List 2. We applied this simplification with 25 new replications. This time, estimates were less biased. The minimum was 1961, the maximum was 2147, the median was 2008, and the mean was 2020.

4 Discussion

The CRC problem is fundamentally a missing data problem. The quantity of interest is n, the population size, but the covariate values $x_{(n_c+i)\cdot}$ are missing for $i = 1, ..., n - n_c$. The missingness is arguably of the worst possible kind, because it is plausible that the missing units are not observed precisely because they are *different* from the observed units, not only in the distribution of covariates but also in how capture probabilities depend on covariates. Traditional conditional likelihood approaches (such as [4]) implicitly assume the *sameness* of the missing data by extending inference on the conditional likelihood to the unobserved data. By contrast, our method isolates and emphasizes the role of any assumption regarding the missing data in the imputation stage, while the nonparametric estimation of Π^* is not "tainted" by missing data assumptions. Whether this bifurcation is of much value remains to be seen with further testing and development of improved nonparametric conditional density estimation methods.

References

[1] Kenneth H. Pollock. Building models of capture-recapture experiments. *Journal of the Royal Statistical Society*, 25(4):253–259, 1976.

[2] Stephen E. Fienberg, Matthew S. Johnson, and Brian W. Junker. Classical multilevel and bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society*, 162(3):383–405, 1999.

[3] Juha M. Alho. Logistic regression in capture-recapture models. *Biometrics*, 46(3):623–635, 1990.

[4] Paul S. F. Yip, Emmy C. Y. Wan, and K. S. Chan. A unified approach for estimating population size in capture-recapture studies with arbitrary removals. *Journal of Agricultural, Biological, and Environmental Statistics*, 6(2):183–194, 2001.

[5] Song Xi Chen and Chris J. LLoyd. Estimation of population size from biased samples using non-parametric binary regression. *Statistica Sinica*, 12, 2002.

[6] Daniel Manrique-Vallier and Stephen E. Fienberg. Population size estimation using individual level mixture models. *Biometrical Journal*, 50(6):1–13, 2008.

[7] Peter Hall, Jeff Racine, and Qi Li. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026, 2004.

[8] Stephen E. Fienberg. The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, 59(3):591, 1972.