Doubly Robust Capture-Recapture Methods for Estimating Population Size

Manjari Das, Edward H. Kennedy, and Nicholas P. Jewell

Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA; Division of Medical Statistics, London School of Hygiene & Tropical Medicine, London, UK; Division of Biostatistics and Epidemiology, University of California, Berkeley, CA

ABSTRACT

Estimation of population size using incomplete lists has a long history across many biological and social sciences. For example, human rights groups often construct partial lists of victims of armed conflicts, to estimate the total number of victims. Earlier statistical methods for this setup often use parametric assumptions, or rely on suboptimal plug-in-type nonparametric estimators; but both approaches can lead to substantial bias, the former via model misspecification and the latter via smoothing. Under an identifying assumption that two lists are conditionally independent given measured covariates, we make several contributions. First, we derive the nonparametric efficiency bound for estimating the capture probability, which indicates the best possible performance of any estimator, and sheds light on the statistical limits of capture-recapture methods. Then we present a new estimator, that has a double robustness property new to capture-recapture, and is near-optimal in a nonasymptotic sense, under relatively mild nonparametric conditions. Next, we give a confidence interval construction method for total population size from generic capture probability estimators, and prove nonasymptotic near-validity. Finally, we apply them to estimate the number of killings and disappearances in Peru during its internal armed conflict between 1980 and 2000.

ARTICLE HISTORY

Received August 2021
Accepted February 2023

KEYWORDS

Abundance estimation; Influence functions; Multiple systems estimation; Nonparametric methods; Semiparametric theory

1. Introduction

Capture-recapture is a study design for estimating population size when only a fraction of the population is observed. This setup arises in ecological abundance, disease prevalence, and casualties in armed conflicts. It has a long history, dating back to Graunt in the 1600s (Hald 2003), who used it to estimate plague prevalence in England. Similarly, in 1802 Laplace estimated the total population of France (Goudie and Goudie 2007), and Petersen (1896) fish abundance. Recently it has been used to estimate the number of pages on the web (Fienberg et al. 1999) and the number of victims in a war (Ball et al. 2003).

The simplest capture-recapture setup, credited to Petersen (1896) and Lincoln (1930), consists of two independent lists with partial captures from the population of interest. There have been many generalizations over time. For our purposes, much of the previous work in capture-recapture can be viewed as falling within one of three streams. The first and oldest stream includes relatively simple data structures, for example, involving no covariate information and relatively few lists (Petersen 1896; Lincoln 1930; Schnabel 1938; Darroch 1958). More recent advances in this stream include Burnham and Overton (1979) and Lee and Chao (1994). A second stream emerged to handle more intricate data structures, for example, complex covariate information to help account for heterogeneity/dependence, largely using model-based approaches (Fienberg 1972; Carothers 1973; Huggins 1989; Alho 1990; Tilling and Sterne 1999; Yip et al. 2001; Pollock 2002; Link 2003). However, the advantages of this second stream typically come at the expense of potentially restrictive parametric modeling assumptions, which when violated would induce bias. A third more recent stream addresses similar data structures as the second, but using more flexible nonparametric tools, for example, local kernel or nonparametric Bayes or spline methods (Chen and Lloyd 2000; Zwane and van der Heijden 2005; Huggins and Hwang 2007, 2011; Stoklosa and Huggins 2012; Yee, Stoklosa, and Huggins 2015; Manrique-Vallier 2016; Kurtz 2014). However, the work in this third stream has relied on typically suboptimal plug-in estimators, which can suffer from nonparametric smoothing bias and slow rates of convergence (van der Laan and Robins 2003; Robins et al. 2008; van der Vaart 2014). We refer to Kurtz (2014) for a more detailed review of this stream.

Our work takes the nonparametric perspective, but uses advances in efficiency theory to characterize optimality and improve simple plug-in estimators (Bickel et al. 1993; van der Laan and Robins 2003; Kennedy 2016). Under an identifying assumption that two lists are conditionally independent given measured covariate information (described in Section 2), we make several contributions.

- In Section 3 we derive the nonparametric efficiency bound for estimating the capture probability, which indicates the best possible performance of any estimator, and sheds light on the statistical limits of capture-recapture methods.
- In Section 4 we present a new doubly robust estimator, and study its finite-sample error, showing it is near-optimal.
in a nonasymptotic sense, under mild nonparametric conditions.

- In Section 5 we give a general method for constructing confidence intervals for population size from generic capture probability estimators, and prove nonasymptotic near-validity.

- In Section 6 we study our methods in simulations, and apply them to estimate the number of killings and disappearances attributable to different groups in Peru during its internal armed conflict between 1980 and 2000.

Another way to frame our overall contribution is as follows. Given two lists that are not independent, two possible ways forward are to: (i) collect more lists and try to model dependence accurately, or (ii) collect covariate information that might help de-correlate the lists. Approach (i) has arguably been the main focus of the capture-recapture literature; there has been substantial important work on approach (ii), but not from the lens of nonparametric efficiency that we use here.

2. Preliminaries

2.1. Setup

Consider a finite population of $n$ individuals, where the size $n$ is unknown and to be estimated. We suppose there are $K$ different lists of individuals from this population, yielding indicators $Y_{ik} \in \{0, 1\}$ of whether individual $i \in \{1, \ldots, n\}$ appeared on list $k \in \{1, \ldots, K\}$. We let $Y_i = (Y_{i1}, \ldots, Y_{iK})^T$ denote the vector indicating list membership (i.e., capture profile) information for individual $i$. For example, in the $K = 2$ case, a profile $Y_i = (1, 0)^T$ would mean that individual $i$ appears on list 1 but not list 2. We consider the case where covariates $X_i \in \mathbb{R}^d$ are available for each individual $i = 1, \ldots, n$. We assume an individual’s chances of appearing on any given lists (and their covariates) do not depend on what happens with any other individuals, and also that the covariate and (conditional) list membership distributions are the same across individuals $i = 1, \ldots, n$. This implies that the random vectors $Z_i = (X_i, Y_i)$ are independent and identically distributed according to some distribution $\mathbb{P}$.

Remark 1. The setup above is commonly referred to as “heterogeneous” (Huggins 1989; Tilling and Sterne 1999; Pollock 2002) since list membership $Y$ can vary with covariates $X$, that is, individuals with different covariates can have different chances of list membership.

Remark 2. In what follows, we use the following standard notation. We let $\mathbb{E}_Q$ denote an expectation under distribution $Q$, and let $\|f\|^2_Q = \int f(x)^2 \, dQ(x)$ denote the corresponding squared $L_2(Q)$ norm; we let $Q_N$ denote the empirical measure under distribution $Q$. Finally, we let $a \lesssim b$ mean $a \leq Cb$ for some universal constant $C$.

If every individual in the population appeared on at least one list (and could be uniquely identified), then the population size would of course be known without error; however, in practice a possibly substantial fraction of individuals does not appear on any list. In other words, there are some individuals with $Y = 0$ that we do not observe. This means that, although the distribution $\mathbb{P}$ governs the capture profiles, we cannot sample from $\mathbb{P}$ directly. Instead, we only see the $N = \sum_{i=1}^n 1(Y_i \neq 0)$ individuals for whom $Y_{ik} = 1$ for some $k$. This is illustrated in Figure 1. Hence, the capture-recapture design is an example of biased sampling (Vardi 1985; Breslow et al. 2000; Qin 2017). In particular, the observed data $Z_i = (X_i, Y_i), i = 1, \ldots, N$ iid draws from a conditional distribution $Q$ defined as

$$Q(Y = y, X = x) \equiv \mathbb{P}(Y = y, X = x \mid Y \neq 0) = \psi^{-1}\mathbb{P}(Y = y, X = x) \mathbb{1}(y \neq 0)$$

where $\psi$ is the (marginal) capture probability defined as

$$\psi = \mathbb{P}(Y \neq 0).$$

Remark 3. Some authors use $N$ to denote the total population size and $n$ for the observed number of captures; in contrast, we use $N$ for the observed number and $n$ for the total size, following the convention of saving upper case for random variables. Note, the observed number of captures $N$ is random since it depends on the random selection indicators, while the total population size $n$ is fixed; specifically $N \sim \text{Bin}(n, \psi)$. Nonetheless, much of our analysis will be conditional on the observed sample size $N$.

Recall, our overall goal is to estimate the total population size $n = N + \sum (Y_i = 0)$. Since $N \sim \text{Bin}(n, \psi)$, the population size can be viewed as a fixed population parameter

$$n = \mathbb{E}(N)/\psi.$$  

Intuitively, the lower the capture probability, the more the observed number $N$ must be inflated to reflect the total population size. The quantity $\mathbb{E}(N)$ in (3) can of course be unbiasedly estimated with the observed number of captures $N$; therefore, estimating population size essentially boils down to estimating the capture probability, which can then be used to inflate the observed $N$ via the estimator

$$\hat{n} = N/\hat{\psi}.$$  

Thus, we turn toward the crucial question of how to efficiently estimate the capture probability (specifically, in the presence of
high-dimensional and/or complex covariates $X$), before coming back to inference about $n$ in Section 5. In the next section we discuss identification of $\psi$ from the distribution $Q$ of which we sample; this will require extra assumptions, since if lists are irreparably dependent we will have no information about those who are unobserved.

**Remark 4.** In what follows, we focus on the case where there are two primary lists of interest, $k = 1, 2$, so that each captured individual contributes a vector $Z_i = (Y_{i1}, Y_{i2}, X_i)$. The captured individuals that do not appear on list 1 or 2 (but do appear on some other list) therefore contribute $Z_i = (0, 0, X_i)$. More discussion on this point follows in the next section.

### 2.2. Identification

As mentioned in the previous section, without additional assumptions, the observed data distribution $Q$ of list membership among those on at least one list is completely uninformative about the capture probability $\psi = P(Y \neq 0)$. A variety of assumptions have been used to identify this and related quantities in previous literature; broadly, there must be some lack of dependence across lists in order to identify and estimate the odds and thus the overall population size. The Lincoln-Petersen estimator (Petersen 1896; Lincoln 1930) assumed independence among the lists which are completely uninformative without loss of generality. We extended it to assume independence across $K > 2$ lists. Fienberg (1972) assumed a log-linear model for the expected number of observations with each capture profile, with one parameter set to zero (typically the highest-order interaction term across lists). You et al. (2021) presented a framework adaptable to various identification assumptions, including log-linear model assumption and independence between two lists conditional on the remaining list(s) for $K > 2$ case.

When one has access to not only list membership but also covariate information, conditional versions of these assumptions can be used (Sekar and Deming 1949; Chao 1987; Tilling and Sterne 1999; Huggins and Hwang 2007). Importantly, this allows heterogeneous capture probabilities that vary across units, thus, relaxing identifying assumptions; this is conceptually similar to how measured confounders are exploited in observational studies for causal inference (Hernan and Robins 2019). Sekar and Deming (1949) studied the conditional case in the discrete and low-dimensional setup; the continuous case has been studied using parametric models by Pledger (2000), Pollock et al. (1990), Tilling and Sterne (1999), Huggins (1989), Chao (1987), and Alho (1990). Burnham and Overton (1979), and Huggins and Hwang (2007) used a jackknife estimator. Analogous to the assumption of Tilling and Sterne (1999) for the two list case, our main identifying assumption for the $K$ list case is that there is a known pair among the $K$ lists which are conditionally independent. Without loss of generality, we order the lists, so the first two are conditionally independent:

**Assumption 1.** $P(Y_1 = 1 \mid X = x, Y_2 = 1) = P(Y_1 = 1 \mid X = x, Y_2 = 0)$.

Under Assumption 1 (e.g., as in Tilling and Sterne 1999), the capture probability $\psi = P(Y \neq 0)$ can be identified from the biased observed data distribution $Q$. Specifically, let $q_1(x) = Q(Y_1 = 1 \mid X = x)$, $q_2(x) = Q(Y_2 = 1 \mid X = x)$, $q_{12}(x) = Q(Y_1 = 1, Y_2 = 1 \mid X = x)$ denote the observational probability (under $Q$) of appearing on list 1, 2, and both, respectively. These probabilities will be referred to as the $q$-probabilities throughout.

**Remark 5.** Note that when there are only $K = 2$ lists, it must be that $q_1(x) + q_2(x) = q_{12}(x) = 1$ since each observed unit must appear on list 1, list 2, or both, according to the sampling distribution $Q$. In general, when $K > 2$ it only holds that $0 \leq q_1(x) + q_2(x) - q_{12}(x) \leq 1$, since some individuals may only appear on lists $j \geq 3$ other than 1 and 2.

**Remark 6.** Note that when there are more than two lists and without loss of generality the conditionally independent list pair is 1 and 2, the remaining lists aid the estimation by potentially increasing the number of observed individuals (as described in Remark 4, our method does not discard individuals not captured in list 1 or 2). This in turn leads to variance reduction, as discussed in Appendix A.2, supplementary materials. Existing methods that use the information of all the lists include log-linear models, Bayesian latent-class models, or conditioning on the remaining lists (Fienberg 1972; Manrique-Vallier 2016; You et al. 2021).

For posterity, we give the identification result for $\psi$ in the following proposition (with a proof given in the Appendix, supplementary materials).

**Proposition 1.** Under Assumption 1 and the positivity condition $Q(q_{12}(x) > 0) = 1$, the conditional and marginal capture probabilities are identified from $Q$ by

$$
\gamma(x) = \frac{q_{12}(x)}{q_1(x)q_2(x)}
$$

$$
\psi = P(Y \neq 0) = \left(\int \gamma(x)^{-1} dQ(x)\right)^{-1}
$$

**Remark 7.** We note that the positivity condition $Q(q_{12}(x) > 0) = 1$ is not unique to our approach—if there is zero overlap, no method should be able to estimate population size well (in the conditional independence model considered here, without further assumptions). This condition is analogous to positivity conditions in causal inference (e.g., Westreich and Cole 2010; Petersen et al. 2012; Kennedy 2019, etc.). Two typical strategies used in that literature are: (i) truncate estimated probabilities (e.g., set $\hat{q}_{12}(x) = 1\%$ whenever a method returns an estimate smaller than 1%), or else (ii) target different less sensitive parameters (with the tradeoff that they may be more difficult to interpret). The latter approach has been used in capture-recapture by Johnord et al. (2019).

In the following sections, we give three main contributions. First we derive the efficiency bound for estimating the capture probability $\psi$ under a nonparametric model that puts no parametric restrictions on the "nuisance" functions ($q_1, q_2, q_{12}$); second, we construct novel estimators that attain the efficiency bound under weak nonparametric conditions (e.g., allowing the use of flexible machine learning tools); and third, we give a
general method for building corresponding confidence intervals for the total population size $n$, given any asymptotically linear estimate $\hat{\psi}$ of the capture probability.

Remark 8. All subsequent results apply to the statistical parameter $\psi$, which we define as the harmonic mean on the right-hand-side of (6). Under the identifying Assumption 1 (and positivity), $\psi$ also represents the capture probability on the left-hand-side of (6), but our statistical results do not require this link, and apply to the harmonic mean in (6) regardless.

Remark 9. We stress that Assumption 1 is agnostic about the presence or absence of additional dependence structure beyond lists 1–2. When such extra structure does exist, our approach would still be valid, but likely not as efficient as a method that correctly assumed and exploited this structure (e.g., the methods referenced at the end of Remark 6). Similarly, when there are more than two lists, it is possible that a better mean squared error in a local minimax sense (van der Vaart 2002b). The following theorem and corollary give the form of this bound and formalize the minimax result. All expectations and variances are under distribution $\mathbb{Q}$ unless noted otherwise.

Theorem 1. Let $g: \mathbb{R} \mapsto \mathbb{R}$ be any function differentiable at $\psi$. The nonparametric efficiency bound for estimation of $g(\psi)$ is given by $\var(g(\psi)) = f_q(\psi)^2 \sigma^2$, where $f_q(\psi)$ is defined in Lemma 1,

$$
\sigma^2 = \mathbb{E} \left( \frac{1}{\gamma(X)} \left[ \frac{1 - \gamma(X)}{\gamma(X)} \var \left( \frac{1 - q_{12}(X)}{q_{12}(X)} \right) \right] + \frac{q_0(X)}{q_{12}(X)} \right).
$$

and $q_0(x) = 1 - q_1(x) - q_2(x) + q_{12}(x)$ is the chance of appearing on neither list 1 nor 2.

The magnitude of the efficiency bound in Theorem 1 is driven by three main factors:

(i) the magnitude of the conditional capture probabilities, 
(ii) the chance of appearing on both lists, and
(iii) the heterogeneity in the conditional capture probabilities.

Remark 10. The efficiency bound for any function $g(\cdot)$ is always proportional to $\sigma^2$, with a scaling $f_q(\psi)^2$ depending on $g$; for example, $f_q(\psi) = -1$ when $g(\psi) = 1/\psi$, and $f_q(\psi) = \psi/(1 - \psi)$ when $g(\psi) = \logit(\psi)$. Therefore, we focus our discussion on the quantity $\sigma^2$.

The dependence on (i) in the bound in Theorem 1 occurs through the term $(1 - \gamma)/\gamma^2$, that is, the odds of capture divided by the capture probability. The dependence on (ii) occurs through the odds $(1 - q_{12})/q_{12}$ as well as the probability ratio $q_0/q_{12}$. The dependence on the heterogeneity (iii) occurs through the term $(1/\gamma)$. Note that the probabilities $\gamma$ and $q_{12}$ in (i) and (ii) are related, but $q_{12}$ can be small even when the capture probability $\gamma$ is not, depending on the size of $q_1$ and $q_2$.

More specifically, all else equal, the variance bound increases with: (i) smaller capture probabilities $\gamma$, (ii) smaller chances of appearing on both lists $q_{12}$, and (iii) greater heterogeneity in the capture probabilities $\gamma$. Therefore, capture probabilities can be
estimated most efficiently when capture is likely, when there is
substantial overlap across lists, and when capture probabilities
are more homogeneous.

**Remark 11.** For \( K = 2 \), the quantity \( q_0(x) \) is exactly zero, but
when \( K > 2 \) it can be positive.

**Remark 12.** In the absence of covariates for \( K = 2 \), the quantity
\( \sigma^2 \) reduces to \( \left( \frac{1}{\hat{\psi}} \right) \left( \frac{1-q_{12}}{q_{12}} \right) \).

Beyond informing factors that yield more efficient capture
probability estimation, Theorem 1 also acts as a local minimax
lower bound, as formalized in the following corollary.

**Corollary 1.** For any estimator \( g(\hat{\psi}) \), it follows that
\[
\inf_{\delta > 0} \liminf_{N \to \infty} \sup_{TV(\tilde{\xi}, \xi) < \delta} \frac{\mathbb{E} \left[ (g(\hat{\psi}) - g(\hat{\psi}))^2 \right]}{f_\delta(\hat{\psi})^2 (\sigma^2 / N)} \geq 1
\]
where \( TV(\tilde{\xi}, \xi) \) is the total variation of \( \tilde{\xi} \) and \( \xi \), \( \hat{\psi} = \psi(\tilde{\xi}) \) and \( \psi = \psi(\xi) \) are the capture probabilities under \( \tilde{\xi} \) and \( \xi \),
and \( f_\delta(\hat{\psi}) \) and \( \sigma^2 = \sigma^2(\xi) \) are defined as in Theorem 1.

Corollary 1 shows that the worst-case mean squared error of
any estimator of \( \psi \), locally near the true \( \psi \), cannot be smaller
than the efficiency bound, asymptotically and after scaling by \( N \).
This local minimax result gives an important benchmark
for efficient estimation of the inverse capture probability: no estimator
can have mean squared error uniformly better than the variance
of the efficient influence function divided by \( N \), without
adding extra assumptions and/or structure to the nonparametric
model we consider.

**Remark 13.** The local minimax result in Corollary 1 holds for
any subconvex loss function \( \ell : \mathbb{R} \to [0, \infty) \) applied to
\( \sqrt{N} (g(\hat{\psi}) - g(\hat{\psi})) \), not just squared error loss \( \ell(t) = t^2 \); the
denominator lower bound in the general case is \( \mathbb{E} (\ell(f_\sigma Z)) \) where
\( Z \sim N(0,1) \) is a standard normal random variable
(van der Vaart 2002b).

Importantly, in the next section we construct estimators that
in each case construct estimators that
can achieve the efficiency bound under weak conditions that
allow for flexible estimation of the \( q \)-probabilities, for example,
using machine learning tools.

## 4. Efficient Estimation

### 4.1. Setup

Recall, we let \( Q_N \) denote the empirical measure under \( Q \), so that
sample averages can be written with the short-hand \( Q_N(f) = Q_N(f(Z)) = \frac{1}{N} \sum_{i=1}^N f(Z_i) \). The simplest estimator of the capture
probability \( \psi \) is just a plug-in
\[
\hat{\psi}_{pl} = \left[ \frac{1}{\hat{\psi}(X)} \right]^{-1}
\]
which replaces unknown quantities in the definition of \( \psi \) with
estimates, that is, by estimating the conditional capture prob-
ability \( \hat{\psi}(X) = \frac{q_{12}(X)}{q_1(X)q_{12}(X)} \) for every unit, and computing the
harmonic mean of the values across the sample. This estimator
has been used relatively extensively in previous work (Darroch
1958; Fienberg 1972; Tilling and Sterne 1999; Huggins and
Hwang 2007). When the \( q \)-probabilities are estimated with
correctly specified parametric models, the plug-in estimator \( \hat{\psi}_{pl} \)
will be \( \sqrt{N} \)-consistent and asymptotically normal under standard
regularity conditions. However, when the covariates contain
any continuous components and/or are high-dimensional, it is
usually very unlikely an analyst would have enough a priori
knowledge to be able to correctly specify a low-dimensional
parametric model, let alone three (one for each \( q \)-probability
nuisance function).

This difficulty of correct model specification suggests trying to
flexibly estimate the \( q \)-probabilities, for example, using logistic
regression with model selection, or lasso, or nonparametric
tools like random forests, neural nets, RKHS regression, etc.
Unfortunately, when the plug-in estimator \( \hat{\psi}_{pl} \) is constructed
from these kinds of data-adaptive methods, it in general loses
the nice properties it has in the parametric setup. Specifically,
without special tuning of particular methods, it in general would
suffer from slower than \( \sqrt{N} \)-convergence rates, and have an
unknown limiting distribution, making it not only inefficient
but also leaving no tractable way to do inference. This defi-
cency of plug-in estimators is by now relatively well-known in
functional estimation problems (van der Laan and Robins 2003;
Chernozhukov et al. 2018; Wu and Yang 2019); however, we
have not seen it highlighted in the capture-recapture setting. (We
show these issues via simulations in Section 6.1.)

Luckily, the plug-in can be improved upon using semipara-
metric efficiency theory (Bickel et al. 1993; van der Vaart 2002a;
van der Laan and Robins 2003; Tsaitis 2006; Kennedy 2016). In
what follows, we present a doubly robust estimator, which can
attain the efficiency bound from the previous section even when
built from data-adaptive regression tools.

### 4.2. Doubly Robust Estimator

As mentioned above, the plug-in estimator (8) has some
important deficiencies in semi- and nonparametric settings.
The plug-in estimator can be debiased by adding an estimate of
the mean of the efficient influence function (Bickel et al. 1993;
vander Vaart 2002a; van der Laan and Robins 2003; Tsaitis 2006;
Kennedy 2016). This leads to our proposed doubly robust estimator
\[
\hat{\psi}_{dr} = Q_N \left[ \frac{1}{\hat{\psi}(X)} \left\{ \frac{Y_1}{\hat{q}_1(X)} + \frac{Y_2}{\hat{q}_2(X)} - \frac{Y_1 Y_2}{\hat{q}_1 \hat{q}_2(X)} \right\} \right]^{-1}
\]
where \( \hat{q}_i \) are estimates of the \( q \)-probabilities (e.g., via regression
predictions).

**Remark 14.** In order to avoid potentially restrictive empiri-
cal process conditions, we estimate \( Q_N \) and \( \hat{q}_i \) from separate
independent samples. Specifically, we estimate the \( q \)-probability
nuisance functions by fitting regressions in a training sample,
independent of a test sample \( Q_N \). With iid data, one can always
obtain such samples by splitting at random in half, or folds. This
yields a loss in efficiency, but that can be fixed by swapping the
samples/folds, computing the estimate on each, and averaging.
This is referred to as cross-fitting, and has been used for example
by Bickel and Ritov (1988), Robins et al. (2008), Zheng and van der Laan (2010), and Chernozhukov et al. (2017). Here we analyze a single split procedure, merely to simplify notation; extending to averages across independent splits is straightforward.

In the following section, we derive finite-sample error bounds and distributional approximations for our doubly robust estimator, which are valid for any sample size.

4.2.1. Nonasymptotic Error Bounds and Approximate Normality

In this section, we provide three main theoretical results regarding error bounds for our proposed method. In particular, we show that our estimator is nearly efficient, doubly robust, and approximately normal. Importantly, we show all these properties hold in finite samples, without resorting to asymptotics.

In the previous section, we derived the efficient influence function, which is the crucial component of the local minimax lower bound we gave in Corollary 1. This corollary shows the minimax optimal estimator has a mean squared error that scales like the variance of the efficient influence function divided by \( N \), so that an optimal estimator would be one that behaves like a sample average of the efficient influence function. Our first result shows that our proposed estimator does in fact behave like an average of the efficient influence function, depending on the size of nuisance error. In what follows, we use \( \phi \) and \( \hat{\phi} \) to denote the efficient influence function \( \phi(Z; \xi) \) and its estimate \( \hat{\phi}(Z; \hat{\xi}) \), respectively.

**Theorem 2.** For any sample size \( N \) and error tolerance \( \delta > 0 \), we have

\[
| (\hat{\psi}_{dr}^{-1} - \psi^{-1}) - \xi_N \phi | \leq \delta
\]

with probability at least

\[
1 - \left( \frac{1}{\delta^2} \right) \mathbb{E} \left( \frac{N}{\hat{\phi} - \phi} \right)^2
\]

where \( \hat{\phi} \) is a second-order error term given by

\[
\hat{\phi} = \int \frac{1}{q_{12}} \left[ \left( q_1 - \hat{q}_1 \right) \left( q_2 - q_2 \right) + \left( q_{12} - \hat{q}_{12} \right) \left( 1 - \frac{1}{\gamma} \right) \right] d\xi
\]

\[
\leq \left( \frac{1}{\epsilon} \right) \left| q_1 - \hat{q}_1 \right| \left| q_2 - q_2 \right| + \left( \frac{1}{\epsilon^2} \right) \left| q_{12} - \hat{q}_{12} \right| \left| \gamma - \hat{\gamma} \right|
\]

with the latter bound on \( \hat{\phi} \) holding as long as \( q_{12} \) and \( \hat{q}_{12} \) are accurate.

**Theorem 2** shows that our proposed estimator is within \( \delta \) of a sample average of the efficient influence function, centered at the true (inverse) capture probability, with high probability, at every sample size. For a given observed number of captures \( N \) and error \( \delta \), this probability depends on two factors: (i) a second-order error term \( \hat{\phi} \), which is driven by the error in estimating the nuisance \( q \)-probabilities; and (ii) the \( L_2 \) error in estimating the efficient influence function itself, divided by \( N \), which also depends on estimation error of the \( q \)-probabilities, but in a weaker way due to the division by \( N \). When \( \hat{\phi} \) goes to 0 as \( N \) increases, the probability above goes to 1 for any fixed \( \delta \). Hence, **Theorem 2** implies usual asymptotic convergence in probability, but it also gives an error bound valid for any \( N \).

For example, if the \( q \)-probabilities are estimated with errors upper bounded by \( c \sqrt{N^{-1/4}} \), then with at least 95% probability, our proposed estimator will be within \( 4c \sqrt{N^{-1/4}} \) of the average efficient influence function. More generally, if \( \mathbb{E}[R_1] \leq 1/\sqrt{N} \), then our proposed estimator will be within \( 1/\sqrt{N} \) (up to constants) of this efficient average, with high probability. For example, if \( q_1, q_2, q_{12} \), and \( \gamma \) belong to Holder classes \( \mathcal{H}(\beta_1), \mathcal{H}(\beta_2), \mathcal{H}(\beta_0) \), and \( \mathcal{H}(\beta_\varphi) \), respectively, where \( \mathcal{H}(s) \) is a Holder class with smoothness index \( s \) (Győrfi et al. 2006; Tsybakov 2008), then a sufficient condition for this kind of result, if the \( q \)-probabilities are estimated at minimax optimal rates, would be that the minimum smoothness is at least half the dimension of the covariates, that is, \( \min(\beta_0, \beta_1, \beta_2, \beta_\varphi) \geq d/2 \). Similarly, if the \( q \)-probabilities were \( s \)-sparse, then a sufficient condition would be that \( s \leq \sqrt{\sqrt{N}/\epsilon} \) with lasso-style methods (Farrell 2015). However, such \( N^{-1/4} \) nuisance errors are only sufficient conditions for \( 1/\sqrt{N} \) capture probability errors; one only needs the remainder error \( R_2 \) to be small enough, which could also be achieved if some combinations of \( q \)-probabilities are estimated well, even if others are not. We give more detail on this phenomenon in our next result.

Namely, we next show that our proposed estimator enjoys a finite-sample multiple robustness phenomenon, never before shown in capture–recapture problems. This phenomenon indicates that the overall estimation error can be small as long as some, but not all, nuisance probabilities are estimated with small error.

**Corollary 2.** Suppose \( (q_{12} \wedge \hat{q}_{12}) \geq \epsilon \). Assume one of the following holds:

1. \( \| q_1 - q_1 \| \vee \| q_{12} - q_{12} \| \leq \xi_N \), or
2. \( \| q_1 - q_1 \| \vee \| \gamma - \hat{\gamma} \| \leq \xi_N \), or
3. \( \| q_2 - q_2 \| \vee \| q_{12} - q_{12} \| \leq \xi_N \), or
4. \( \| q_2 - q_2 \| \vee \| \gamma - \hat{\gamma} \| \leq \xi_N \).

Then for a constant \( C \) independent of \( N \), \( | (\hat{\psi}_{dr}^{-1} - \psi^{-1}) - \xi_N \phi | \leq \delta \) with probability at least

\[
1 - \left( \frac{C}{\delta^2} \right) \left( \frac{1}{N} + \frac{1}{\sqrt{N}} \right)
\]

As a consequence of **Corollary 2**, the proposed estimator is doubly robust, that is, if either estimator of \( q_1 \) or \( q_2 \) has small error, and either estimator of \( q_{12} \) and \( \gamma \) has small error, then the overall error of our proposed estimator (given by \( \hat{R}_2 \)) will be just as small, up to constants, even if the other estimators have large errors or are misspecified. (We note that, although this kind of robustness is sometimes called multiple robustness (Vansteelandt et al. 2008), we use the term doubly robust since the error structure is still second-order, that is, involving products of errors, albeit with more terms). This property is very useful when one of the lists is difficult to estimate, for example, due to high-dimensional covariates, or \( q \)-probabilities that are complex functions of continuous covariates.

**Remark 15.** Note that when there are only \( K = 2 \) lists, then we have the relation \( q_1(x) + q_2(x) - q_{12}(x) = 1 \) for all \( x \).
Hence, small bias for any two estimators of $q_{12}$, $q_1$, and $q_2$ automatically implies small bias for the third. One might expect then that double robustness does not arise in the $K = 2$ list setting; however, this is not quite true. To see why, note that it could be possible to estimate $\gamma$ with small error, for example, even when some of the $q$-probability estimators are misspecified. These and other issues related to estimation of the conditional capture probability $\gamma$ will be important to explore in future work.

When the remainder error $\tilde{R}_2$ is sufficiently small, Theorem 2 and Corollary 2 tell us that we can approximate $\hat{\psi}_{dr}^{-1}$ with a sample average of the efficient influence function. For the purposes of inference, this suggests a confidence interval of the form

$$\hat{C}I = [\hat{\psi}_{dr}^{-1} \pm z_{1-\alpha/2}\hat{\sigma}/\sqrt{N}],$$

(10)

where $\hat{\sigma}$ is a variance term defined in Theorem 3. In the next Berry-Esseen-type result, we exploit this closeness with a sample average and further show that our proposed estimator, properly scaled, is approximately Gaussian. This will show that the above confidence interval gives nearly valid finite-sample coverage guarantees.

**Theorem 3.** Let $\hat{\sigma}^2 = \sqrt{\kappa}(\hat{\phi})$ be the unbiased empirical variance of the estimated efficient influence function. Then $\hat{\psi}_{dr}^{-1} - \psi^{-1}$ follows an approximately Gaussian distribution, with the difference in cumulative distribution functions uniformly bounded above by

$$\left| \mathbb{P}\left( \frac{\hat{\psi}_{dr}^{-1} - \psi^{-1}}{\hat{\sigma}/\sqrt{N}} \leq t \right) - \Phi(t) \right| \leq \frac{C}{\sqrt{N}} \mathbb{E}\left( \frac{\rho}{\hat{\sigma}^2} \right) + \frac{1}{\sqrt{2\pi}} \mathbb{E}\left( \sqrt{N} \mathbb{E}\left( \frac{\tilde{R}_3}{\hat{\sigma}} \right) \right) + \left| t \mathbb{E}\left( \frac{\hat{\sigma} - \sigma}{\sigma} \right)^2 \right|$$

(11)

where $\tilde{\sigma} = \text{var}(\hat{\phi}|Z^n)$, $\rho = \mathbb{E}(\hat{\phi} - \mathbb{E}\hat{\phi}|Z^n)$ and $C < 1/2$ is the Berry-Esseen constant.

The above result shows that the estimation error scaled by $\hat{\sigma}/\sqrt{N}$ is approximately standard normal. The first term on the right-hand side of (11) is the usual Berry-Esseen bound. The second term captures the effect of the nuisance estimation error $\tilde{R}_2$. The third term is the estimation error in the variance. Since $\mathbb{E}(|\hat{\sigma} - \sigma|)$ is bounded above by $C\kappa^{N^{-1/2}}$ (proof in the Appendix, supplementary materials), the overall error in the Gaussian approximation is driven by the second term, involving nuisance error $\tilde{R}_2$. This will be the main driver of whether the interval has approximately correct coverage. We note that the above theorem implies convergence in distribution whenever $\mathbb{E}|\tilde{R}_2| = o(1/\sqrt{N})$ (which can hold for a wide variety of flexible nonparametric estimators of the $q$-probabilities, as discussed after Theorem 2), but in addition gives a more precise error bound that holds for any finite sample size.

Note that Theorem 3 immediately implies that the error in coverage $\left| \mathbb{P}(\hat{C}I \ni \psi^{-1}) - (1 - \alpha) \right|$ for the proposed confidence interval defined in (10) is no more than twice the error bound on the right-hand side of (11), with $t = z_{1-\alpha/2}$. Further, a Berry-Esseen-style bound similar to that of Theorem 3 (along with subsequent coverage guarantees and corollaries) can be obtained for any function $g(\cdot)$ of $\hat{\psi}_{dr}^{-1}$ satisfying the conditions from Friedrich (1989). This implies the same kind of coverage guarantees for $\psi$, for example, using the confidence interval

$$\hat{\psi}_{dr} \pm z_{1-\alpha/2}\hat{\sigma}/\sqrt{N}$$

which can be motivated via the delta method. The error in the coverage of this estimated interval is twice the bound in Theorem 3, modulo some extra dependence on $g$.

Importantly, the unbiased empirical variance $\hat{\sigma}^2$ is a consistent estimator of the efficiency bound $\sigma^2 = \text{var}(\phi)$ in the sense that $\mathbb{E}|\sigma - \hat{\sigma}| \leq \mathbb{E}|\phi - \hat{\phi}| + N^{-1/2}$. This shows the crucial result that our estimator is approximately minimax optimal in the sense of Corollary 1, if the nuisance error is small enough.

A natural consequence of the above theorem is the following corollary, which presents a simple bound on the error of the normal approximation, under some natural conditions on the nuisance error $\tilde{R}_2$ and variance.

**Corollary 3.** Assume $\tilde{\sigma} \geq 1$, $\mathbb{E}|\tilde{R}_3| \leq N^{-2\beta}$ and $\alpha > \delta$ for some $\delta > 0$. Then the coverage error for the proposed $(1 - \alpha)$ confidence interval defined in (10) is upper bounded by

$$\left| \mathbb{P}(\hat{C}I \ni \psi^{-1}) - (1 - \alpha) \right| \leq N^{(1-4\beta)/2} + \frac{1}{\sqrt{N}}.$$

Therefore, if $\beta > 1/4$ there exists some sample size $N_\epsilon$ at which the coverage error is never more than $\epsilon$, for any $N > N_\epsilon$.

Since this corollary is a special case of Theorem 3, mainly aimed at presenting the result in a simple form, we refer to the above discussion for more details. However, we note that the condition that $\mathbb{E}|\tilde{R}_3| \leq N^{-2\beta}$ would hold for example if the $q$-probabilities were estimated optimally when contained in Holder classes with smoothness index $s$, where $\beta = s/(s+\tau)$ (under some conditions on sparsity, as discussed after Theorem 2). Then $\beta > 1/4$ would mean $s > d/2$, aligning with our earlier results.

In this section, we have given finite-sample error bounds and distributional approximations for our proposed estimator, valid for any sample size, allowing accurate estimation and approximately valid confidence guarantees, even in complex nonparametric models where the $q$-probabilities are estimated with flexible machine learning tools. Next we consider a slightly modified version of the estimator, which could further improve finite-sample properties.

**Remark 16.** We have focused on the setting where the goal is to estimate the overall capture probability and/or population size. In practice it may also be of interest to estimate conditional versions of these, for example, the conditional capture probability $\psi(v) = \mathbb{P}(Y \neq 0 | V = v)$ for some subset of covariates $V \subseteq X$. When $V$ is low-dimensional and discrete, one can simply use the same estimator (9) except averaging only among those with $V = v$. Note that when $V$ is high-dimensional or has continuous components, there would be very few if any units with $V = v$ to average over; efficient estimation of the conditional capture probability in this setting is an open problem, which we pursue in future work.
4.3. Targeted Maximum Likelihood Estimator

The proposed doubly robust estimator (9) is close in a finite-sample sense to an optimal sample average, and possesses crucial double robustness properties. However, it is possible this estimator may not respect the bounds on the parameter space; for example, \( \hat{\psi}_{dr} \) may fall outside \([0, 1]\) if some of the estimates of the \( q \)-probabilities are small. A simple fix is to truncate the estimator \( \hat{\psi}_{dr} \) to always lie in \([0, 1]\). Here we discuss an alternative approach using targeted maximum likelihood estimation (TMLE) (van der Laan and Rubin 2006; van der Laan and Rose 2011), which is an iterative procedure that fluctuates nuisance estimates so that a plug-in estimator built from them also approximately solves an efficient influence function estimating equation. TMLE thus leads to estimators that are asymptotically equivalent to one-step bias-corrected estimators, but which could bring some finite-sample advantages.

In Appendix B, supplementary materials, we present an algorithm (Algorithm 1) detailing the computation of a TMLE for \( \psi \). At a high level, the procedure involves bias correction via iterative updating of initial nuisance estimates, based on quantities called clever covariates in the TMLE literature. Interestingly, in addition to being somewhat more computationally intensive, TMLE estimators are not sample averages like our main proposed estimator from the previous section; this makes it less clear how to derive finite-sample error bounds. Since the estimates \( \hat{q}^n(x) \) obtained after convergence satisfy \( \hat{Q}_n(\phi(Z, \hat{Q}^n)) \approx 0 \), the asymptotic behavior matches the doubly robust estimator in (9), but for describing finite-sample behavior we resort to simulations, detailed in Section 6.1.

5. Inference for Population Size

In the previous section, we gave doubly robust estimators for the capture probability and studied finite-sample properties. In this section, we give a crucial result that shows how to obtain an approximate confidence interval for the population size, given a generic initial estimator of the (inverse) capture probability. Importantly, our results only require this initial estimator to be weakly approximated by a sample average, and otherwise are completely agnostic to how the capture probability is estimated. This appears in stark contrast to most of the literature on this topic, where the inferential procedures are very closely tied to specific model assumptions and estimator constructions.

This main inferential result is given in the following theorem.

**Theorem 4.** Suppose we are given an initial estimator \( \hat{\psi} \) that satisfies

\[
\hat{\psi}^{-1} - \psi^{-1} = \hat{Q}_n(\hat{\phi}) - \int \hat{\phi}(z) d\hat{Q}(z) + \hat{R}_2
\]

for \( \phi \) a generic influence function with mean zero and \( \hat{R}_2 \) an error term. Let \( \hat{\tau}^2 = \hat{\psi} \hat{z}^2 + \frac{1}{\hat{\psi}} \) and \( \tau^2 = \psi \hat{z}^2 + \frac{1}{\psi} (\psi \hat{R}_2 + 1)^2 \), where \( \hat{z}^2 = \text{var}(\hat{\phi}) \) is the unbiased empirical variance of the estimated influence function and \( \hat{z} = \text{var}(\hat{\psi} \mid Z^n) \) the true conditional variance. Then the \((1-\alpha)\) confidence interval given by

\[
\hat{C}_{\tau} = \left[ \hat{n} \pm z_{\alpha/2} \hat{\tau} \sqrt{\hat{n}} \right]
\]

has coverage error upper bounded as

\[
\begin{align*}
|\mathbb{P}(\hat{C}_{\tau} \ni n) - (1-\alpha)| &\leq \frac{2C}{\sqrt{n}} \mathbb{E} \left( \frac{\rho}{\sqrt{F}} \right) \\
&\quad + \sqrt{\frac{2}{\pi}} \left[ \sqrt{n} \psi \mathbb{E} \left( \frac{\hat{R}_2}{\sqrt{F}} \right) \right] + |z_{\alpha/2}| \mathbb{E} \left( \left| \frac{\hat{\tau} \sqrt{n}}{\sqrt{F}} - 1 \right| \right)
\end{align*}
\]

where \( C \) is the Berry-Esseen constant and

\[
\rho = \mathbb{E}\left[ \mathbb{1}(Y \neq 0) \left( \hat{\psi} - \hat{\psi} \hat{Q} \right) + \mathbb{1}(Y \neq 0) - \psi \right] \hat{R}_2
\]

\[
+ \psi^{-1} \left( \mathbb{1}(Y \neq 0) - \psi \right)^3 \mathbb{E} \left( \left| Z^n \right| \right)
\]

**Theorem 4** gives an upper bound on how much the coverage \( \mathbb{P}(\hat{C}_{\tau} \ni n) \) of our interval \( \frac{\hat{n}}{\psi} \pm z_{\alpha/2} \sqrt{\left( \frac{\psi \hat{z}^2 + \frac{1}{\psi}}{\psi} \right) \hat{n}} \) can deviate from its nominal \((1-\alpha)\) level. Before describing the coverage guarantee, we first describe the proposed interval. The length of this interval is driven by three factors: (i) the estimated odds of not being captured \( (1 - \hat{\psi}) / \hat{\psi} \), (ii) the variance of the inverse capture probability estimator \( \hat{\tau}^2 \), and (iii) the sample size \( N \). As one would expect, higher odds of capture yield more precise inference about population size, all else equal, as does more efficient estimation of \( \hat{\psi} \). Specifically, the length of the interval shrinks to zero when the capture probability is very large, regardless of the sample size \( N \). Also note that even if \( \hat{\psi} \) were known, one would still have an interval of the form \( \frac{\hat{n}}{\hat{\psi}} \pm z_{\alpha/2} \sqrt{\left( \psi \hat{z}^2 + \frac{1}{\psi} \right) \hat{\psi}} \) based on the fact that \( \hat{n} = N / \hat{n} \) is approximately normal. Although sample size \( N \) appears in the numerator of the interval width (contrary to standard intervals), it only appears through its square root, showing that in an asymptotic regime where \( N \to \infty \), the width still grows at a slower rate than the sample size. Intuitively, this interval takes \( \hat{n} = N / \hat{\psi} \) and multiplies by \( 1 \pm z_{\alpha/2} \sqrt{\hat{n}} \), which does tend to zero as sample size \( N \) grows.

**Remark 17.** For \( K = 2 \) lists and in the absence of covariates, the confidence interval reduces to \( \hat{n} \pm z_{\alpha/2} \sqrt{\frac{\hat{n}(1 - \hat{\psi})}{\hat{\psi} q_{12}}} \), which approximately resembles the Wald-type confidence interval for the Lincoln–Petersen estimator (Evans et al. 1996).

Now we describe the coverage guarantee of **Theorem 4**. Importantly, the bound on the coverage error depends on a number of factors, that is, the sum of the three terms in (13). Under typical boundedness assumptions, the first and third terms would be of smaller order, and the second term would dominate. This second term is driven by the size of \( \hat{R}_2 \) in terms of its mean absolute value, that is, how well the initial estimator \( \hat{\psi} \) is approximated by a sample average. If \( \hat{R}_2 \) is not substantially smaller than \( 1 / \sqrt{\hat{n}} \), then the confidence interval would not be guaranteed to cover the true population size \( n \) at its nominal level. This points to the importance of efficient estimation of \( \psi \); for example, as shown in the previous section, our proposed estimator \( \psi_{dr} \) can be approximated by a sample average up to smaller than \( 1 / \sqrt{\hat{n}} \) error, even in a nonparametric model when \( q \)-probabilities are estimated flexibly.

**Remark 18.** A unique feature of **Theorem 4** is that it is valid for any estimator approximated by a sample average, regardless of
what underlying identification or estimation assumptions were used in its construction. This means if another analyst did not believe the independent lists condition in Assumption 1, and instead constructed an estimate of the capture probability under a different identifying assumption, they could also use the above theorem to construct a confidence interval and assess its finite-sample coverage.

A natural consequence of Theorem 4 is the following corollary, which parallels Corollary 3 in giving a simple bound on normal approximation error, under natural conditions.

**Corollary 4.** Assume \( \hat{\tau} \gtrsim 1, \mathbb{E}[R_2] \lesssim N^{-2\beta} \) and \( \alpha > \delta \) for some \( \delta > 0 \). Then the coverage error for the proposed \((1 - \alpha)\) confidence interval defined in (10) is upper bounded by

\[
\left| \mathbb{P}(\hat{C}_n \ni n) - (1 - \alpha) \right| \lesssim n^{(1-4\beta)/2} + \frac{1}{\sqrt{n}}.
\]

Therefore, if \( \beta > 1/4 \) there exists some population size \( n_e \) at which the coverage error is never more than \( \epsilon \), for any \( n > n_e \).

Since the result in Corollary 4 is similar to that of Corollary 3, we refer there for related discussion. The main point is that, as long as our initial estimator is well-approximated by a sample average, no matter how it was constructed or what assumptions it relies on, our proposed confidence interval (12) will be approximately valid.

### 6. Simulation and Application

So far we have proposed doubly robust estimators for the capture probability, and a general approach for constructing confidence intervals for the total population size, all with nonasymptotic error guarantees. Here we study our methods in simulated data, and apply them to estimate the number of killings in the internal armed conflict in Peru during 1980–2000. The code used to generate the results is available on GitHub at mqnjgrid/capture_recapture.

#### 6.1. Simulation

Here we use simulations similar to Tilling and Sterne (1999), taking \( n = 5000 \) samples from \( X \sim \text{Uniform}(2, 3) \), \( \mathbb{P}(Y_1 = 1 \mid X = x) = \expit(a + 0.4x) \) and \( \mathbb{P}(Y_2 = 1 \mid X = x) = \expit(a + 0.3x) \) where \( a \) takes values \(-2.513, -0.66\) to ensure that the capture probability \( \psi \) takes values \( 0.3, 0.8 \), respectively. This gives sample sizes \( N \) approximately equal to \( \{1500, 4000\} \). Recall that under \( \mathbb{P} \), list membership \( Y_1 \) and \( Y_2 \) are conditionally independent, so the conditional capture probability is \( \psi(x) = 1 - (1 - \expit(a + 0.4x))(1 - \expit(a + 0.3x)) \) and \( q \)-probabilities are equal to \( q_j(x) = \mathbb{P}(Y_j = 1 \mid X = x)/\psi(x) \).

We construct estimates of the \( q \)-probabilities via \( \hat{q}_j(x) = \expit[\logit(q_j(x))] + \epsilon_i \), where we simulate the errors in estimation by \( \epsilon \sim \mathcal{N}(0, \sigma^2) \), and truncate the probabilities at 0.01, as discussed in Remark 7. This allows us to carefully control the error of the \( q \)-probability estimators; since the root mean squared error scales like \( n^{-\alpha} \), this can be viewed as the rate of convergence. We run 500 simulations for each \( \alpha \in \{0.1, 0.2, 0.25, 0.3, 0.4, 0.5\} \). Note \( \alpha \) values 0.5 and 0.25 correspond to the parametric \( n^{-1/2} \) and nonparametric \( n^{-1/4} \) rates, respectively. Figure 2 shows the estimated bias and the root mean square error (RMSE) of \( \hat{\psi} \), along with the coverage proportion for the confidence interval of the total population size.

**Remark 19.** For plug-in estimators, there is no well-defined variance formula (this is a main motivation for our doubly robust construction). Therefore, to construct confidence intervals with the plug-in estimator, we used the estimated variance of the doubly robust estimator.

![Figure 2](image-url). Estimated bias, RMSE, and population size coverage, for simulated data with population size \( n = 5000 \), across true capture probability \( \psi \in \{0.8, 0.3\} \), \( q \)-probability error rate \( n^{-\alpha} \) for \( \alpha \in \{0.1, 0.5\} \), and for three different estimators.
Overall, the simulations illustrate the phenomena expected from our theoretical results: when the \( q \)-probabilities are estimated with low error (i.e., \( \alpha \) large), all the methods do well, whereas when the \( q \)-probabilities are difficult to estimate (i.e., \( \alpha \) smaller) the proposed methods do substantially better in terms of bias, error, and coverage. For example, when the true capture probability is 50\%, the simple plug-in estimator gives substantial bias as soon as \( \alpha < 0.4 \) (i.e., when the \( q \)-probabilities are estimated at slower than \( n^{-2/5} \) rates). However, the bias of the proposed doubly robust estimator is relatively unaffected until \( \alpha < 0.2 \), with the TMLE somewhere in between. The story is similar for the RMSE, which is largely driven by the bias in this problem. The coverage is approximately at the nominal 95\% level as soon as \( \alpha > 0.2 \) (i.e., when the \( q \)-probabilities are estimated at faster than \( n^{-1/5} \) rates), whereas the plug-in estimator substantially under-covers (e.g., nearly zero at \( \alpha = 0.2 \)) until \( \alpha \geq 0.4 \). Using simulated data with capture probability 0.5, one will get results similar to those of \( \psi = 0.3 \). We note that when population size or capture probability is small (e.g., capture probability substantially less than 50\%), estimation becomes more challenging and the story is less clear about which method does better. For reference, results for population sizes from \( n = 200 \) to \( n = 1000 \) (in the \( \alpha = 0.25 \) case) are given in the Appendix in Figure 6, supplementary materials.

**Remark 20.** In a follow-up paper (Das and Kennedy 2021) we conducted a simulation study where \( q \)-probabilities were estimated using random forests and logistic regression, rather than via simulation. We reproduce a plot of these results in the Appendix in Figure 7, supplementary materials.

**6.2. Data Analysis**

We apply our proposed methods to estimate the number of killings and disappearances attributable to different groups in Peru during its internal armed conflict between 1980 and 2000. We use data collected by the Truth and Reconciliation commission of Peru (Ball et al. 2003), as well as detailed geographic information, following Rendón (2019).

There is an ongoing debate regarding the total number of killings and disappearances in the conflict, as well as about which groups are most responsible, for example, the PCP-Shining Path versus the State or other groups. Ball et al. (2003) estimated approximately 69,000 total killings and disappearances, finding the Shining Path responsible for the majority. In contrast, Rendón (2019) estimated approximately 48,000 killings and disappearances, with the State responsible for the majority, though many geographic strata were excluded. Most recently, Manrique-Vallier et al. (2019) included a newly available list and estimated approximately 58,000–65,000 killings and disappearances, depending on choices of priors, with the Shining Path responsible for the majority.

Before describing our specific approach, we first describe the data and give some summary statistics. As explained in Ball et al. (2003), the data come from a few main sources: the Truth and Reconciliation Commission (CVR), the Public Defender Office (DP), and 4–5 other human rights groups and NGOs (ODH). We use the CVR as our first list and construct the second list by combining the remaining lists, that is, DP and ODH, since they have similar demographics. The data contains identifiers of people who have been killed or disappeared, as well as which of the source lists they appeared on, and covariates including age, gender, and geographic location of the killing or disappearance (measured via 58 geographic strata as in Ball et al. (2003), as well as bivariate latitude/longitude as in Rendón 2019). To avoid missing completely at random assumptions, we also included missingness indicators for victims with missing age (28% missing), gender (<1% missing), or location (11% missing). The list of all the covariates is available in Appendix D, supplementary materials. The total number of killings and disappearances across all lists was 24,692. Importantly, the lists capture different demographics, which points to the necessity of relaxing classical marginal independence via the conditional independence in Assumption 1. For example, the CVR list mostly includes victims who were killed, while the DP and ODH lists mostly include victims who disappeared, as shown in the Appendix in Figure 8, supplementary materials. Similarly, geographic diversity varies across lists, as shown in Figure 3. For example, almost 60\% of Shining Path victims in the DP and ODH lists come from two smaller districts (Chungui and Luis Carranzo) of Ayacucho, while in the CVR list the Shining Path victims are more uniformly spread across the country. More details on the data are available in Appendix D, supplementary materials.

Now we move to our analysis. Our goal was to estimate the number of killings and disappearances attributable to the State and Shining Path, as well as those that were not identified as either. We used our proposed doubly robust estimator (9) with 5-fold cross-fitting, and we estimated the \( q \)-probabilities via random forests (using the ranger package in R). We truncated all \( q \)-probability estimates at 0.01. Figure 4 shows the estimated number of killings and disappearances along with 95\% confidence intervals obtained using the interval (12). To estimate numbers of killings and disappearances by group, we used the conditional averaging approach described in Remark 16. We estimate the total number of killings and disappearances across groups to be 68,874 (95\% CI: 58,543–79,204), close to the estimates in Ball et al. (2003) and the diffuse prior-based estimate in Manrique-Vallier and Ball (2019) (which used an additional list). This suggests that using covariate information with the two
aggregated lists is roughly as informative as the multivariate lists used in previous work. Overall we find the State responsible for more disappearances, and Shining Path responsible for more killings; however, we estimate the number of killings and disappearances by unidentified perpetrators to be larger than that for either group. In terms of the overall killings and disappearances, the estimate for the State are higher compared to the estimate for the Shining Path. We present some more details of the analysis and a location wise estimate comparison for the State and the Shining Path in Appendix D, supplementary materials.

7. Discussion

In this article, we study estimation of population size and capture probability where two lists are conditionally independent given measured covariates. We make four main contributions. First, we derive the nonparametric efficiency bound for estimating the capture probability, which indicates the best possible performance of any estimator, in a local asymptotic minimax sense. This kind of lower bound result has not appeared in the literature, even in simple settings without covariates. Second, we present a new doubly robust estimator, and study its finite-sample properties; in addition to double robustness, we show that it is near-optimal in a nonasymptotic sense, under nonparametric conditions. Third, we give a method for constructing confidence intervals for population size from generic capture probability estimators, and prove nonasymptotic near-validity. Fourth, we study our methods in simulations, and apply them to estimate the number of killings and disappearances attributable to different groups in Peru during its internal armed conflict between 1980 and 2000.

Many extensions are possible. For example, instead of assuming a known pair of lists are conditionally independent given covariates, one could use sensitivity analysis and/or partial identification. For example, one could assume that a pair of lists is only nearly conditionally independent, up to some deviation \( \delta \), and estimate bounds on the capture probability and population size accordingly. This relies on weaker assumptions, with the tradeoff of yielding less precise inferences. Another extension would be to flexibly estimate conditional capture probability or population size, given a continuous covariate such as age or time. For example, for the internal armed conflict in Peru, one might estimate the number of victims by age. This would require a nontrivial extension of the current methods, as described in Remark 16. It will also be important to explore efficient estimation in settings with both complex covariates and multivariate list structure, for example, using conditional no-interaction assumptions. For example, one may have multiple lists, where the list dependence between a particular pair is not modified by being on another list.

Supplementary Materials

In the supplement, we provide the proofs for all results presented in this paper along with additional simulation results and details about the real data.

Acknowledgments

The authors thank Mark van der Laan, and Robin Mejia for very helpful discussions.

Funding

Manjari Das and Edward Kennedy gratefully acknowledge support from the Berkman Faculty Development Fund, and Edward Kennedy from NSF Grant DMS1810979.

ORCID

Manjari Das http://orcid.org/0000-0002-6781-6368
Edward H. Kennedy http://orcid.org/0000-0002-0227-158X
Nicholas P. Jewell http://orcid.org/0000-0002-8637-9132

References

Alho, J. M. (1990), “Logistic Regression in Capture-Recapture Models,” Biometrics, 46, 623–635. [1309,1311]
Ball, P., Asher, J., Sulmont, D., and Manrique, D. (2003), How Many Peruvians have Died, Washington, DC: American Association for the Advancement of Science. [1309,1318]
Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993), Efficient and Adaptive Estimation for Semi-parametric Models (Vol. 4), Baltimore: Johns Hopkins University Press. [1309,1312,1313]
Bickel, P. J., and Ritov, Y. (1988), “Estimating Integrated Squared Density Derivatives: Sharp Best Order of Convergence Estimates,” Sankhyā: The Indian Journal of Statistics, Series A, 50, 381–393. [1314]
Breslow, N. E., Robins, J. M., and Wellner, J. A. (2000), “On the Semi-Parametric Efficiency of Logistic Regression under Case-Control Sampling,” Bernoulli, 6, 447–455. [1310]
Burnham, K. P., and Overton, W. S. (1979), “Robust Estimation of Population Size When Capture Probabilities Vary Among Animals,” Ecology, 60, 927–936. [1309,1311]
Carothers, A. (1973), “The Effects of Unequal Catchability on Jolly-Seber Estimates,” Biometrics, 29, 79–100. [1309]
Chao, A. (1987), "Estimating the Population Size for Capture-Recapture Data with Unequal Catchability," Biometrics, 43, 783–791. [1311]
Chen, S. X., and Lloyd, C. J. (2000), "A Nonparametric Approach to the Analysis of Two-Stage Mark-Recapture Experiments," Biometrika, 87, 633–649. [1309]
Chernozhukov, V., Chetverikov, D., Demirer, M., Dufo, E., Hansen, C., and Newey, W. (2017), "Double/Debiased/Neyman Machine Learning of Treatment Effects," American Economic Review, 107, 261–265. [1314]
Chernozhukov, V., Chetverikov, D., Demirer, M., Dufo, E., Hansen, C., Newey, W., and Robins, J. (2018), "Double/Debiased Machine Learning for Treatment and Structural Parameters," The Econometrics Journal, 21, C1–C68. [1313]
Darroch, J. N. (1958), "The Multiple-Recapture Census, I. Estimation of a Closed Population," Biometrika, 45, 343–349. [1309,1311,1313]
Das, M., and Kennedy, E. H. (2021), "drpop: Efficient and Doubly Robust Population Size Estimation in R," arXiv preprint arXiv:2111.07191. [1318]
Evans, M. A., Kim, H.-M., and O’Brien, T. E. (1996), "An Application of Profile-Likelihood based Confidence Interval to Capture-Recapture Estimators," Journal of Agricultural, Biological, and Environmental Statistics, 1, 131–140. [1316]
Farrell, M. H. (2015), "Robust Inference on Average Treatment Effects with Possibly more Covariates than Observations," Journal of Econometrics, 189, 1–23. [1314]
Fienberg, S. E. (1972), "The Multiple Recapture Census for Closed Populations and Incomplete 2k Contingency Tables," Biometrika, 59, 591–603. [1309,1311,1312,1313]
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, 162, 383–405. [1309]
Friedrich, K. O. (1989), "A Berry-Esseen Bound for Functions of Independent Random Variables," The Annals of Statistics, 17, 170–183. [1315]
Giné, E., and Latała, R. (2002), "Random Sum of Independent Random Vectors," Probability Theory and Related Fields, 122, 393–418. [1316]
Giraud, I., Guedj, B., and Huet, S. (2016), "A Theoretical Analysis of the Benjamini-Hochberg Procedure," Biometrics, 72, 1002–1010. [1317]
Görfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006), Theory of Nonparametric Regression, New York: Springer. [1309]
Hald, A. (2003), A History of Probability and Statistics and their Applications before 1750, New York: Springer. [1309]
Hernan, M. A., and Robins, J. M. (2019), Causal Inference, Boca Raton, FL: CRC. [1311]
Huggins, R. (1989), "On the Statistical Analysis of Capture Experiments," Biometrika, 76, 133–140. [1309,1310,1311]
Huggins, R., and Hwang, W.-H. (2007), "Non-Parametric Estimation of Population Size from Capture–Recapture Data when the Capture Probability depends on a Covariate," Journal of the Royal Statistical Society, Series C, 56, 429–443. [1309,1311,1313]
——— (2011), "A Review of the Use of Conditional Likelihood in Capture-Recapture Experiments," International Statistical Review, 79, 385–400. [1309]
Johns, E. J., Lum, K., and Manrique-Vallier, D. (2019), "Low-Risk Population Size Estimates in the Presence of Capture Heterogeneity," Biometrika, 106, 197–210. [1311]
Kennedy, E. H. (2016), "Semiparametric Theory and Empirical Processes in Causal Inference," in Statistical Causal Inferences and their Applications in Public Health Research, eds. H. He, P. Wu, D.-G. (Din) Chen, pp. 141–167, Cham: Springer. [1309,1312,1313]
Lee, S.-M., and Chao, A. (1994), "Estimating Population Size via Sample Coverage for Closed Capture-Recapture Models," Biometrics, 50, 88–97. [1309]
Lincoln, F. C. (1930), Calculating Waterfowl Abundance on the Basis of Banding Returns, Number 118, US Department of Agriculture. [1309,1311]
Link, W. A. (2003), "Nonidentifiability of Population Size from Capture-Recapture Data with Heterogeneous Detection Probabilities," Biometrics, 59, 1123–1130. [1309]
Manrique-Vallier, D. (2016), "Bayesian Population Size Estimation using Dirichlet Process Mixtures," Biometrics, 72, 1246–1254. [1309,1311]
Manrique-Vallier, D., and Ball, P. (2019), "Reality and Risk: A Retraction of S. Rendón’s Analysis of the Peruvian Truth and Reconciliation Commission’s Conflict Mortality Study," Research & Politics, 6, 2053168019835628. [1318]
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. [1318]
Petersen, C. G. J. (1896), "The Yearly Immigration of Young Plaice in the Limfjord from the German Sea," Report of the Danish Biological Station, 6, 1–48. [1309,1311]
Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y., and Van Der Laan, M. J. (2012), "Diagnosing and Responding to Violations in the Positivity Assumption," Statistical Methods in Medical Research, 21, 31–54. [1311]
Pledger, S. (2000), "Unified Maximum Likelihood Estimates for Closed Capture–Recapture Models Using Mixtures," Biometrics, 56, 434–442. [1311]
Pollock, K. H. (2002), "The Use of Auxiliary Variables in Capture-Recapture Modelling: An Overview," Journal of Applied Statistics, 29, 85–102. [1309,1310]
Pollock, K. H., Nichols, J. D., Brownie, C., and Hines, J. E. (1990), "Statistical Inference for Capture-Recapture Experiments," Wildlife Monographs, 107, 3–97. [1311]
Qin, J. (2017), Biased Sampling. Over-identified Parameter Problems and Beyond, Singapore: Springer. [1310]
Rendón, S. (2019), "Capturing Correctly: A Reanalysis of the Indirect Capture–Recapture Methods in the Peruvian Truth and Reconciliation Commission," Research & Politics, 6, 2053168018820375. [1318]
Robins, J., Li, L., Tchetgen, E., and van der Vaart, A. (2008), "Higher Order Influence Functions and Minimax Estimation of Nonlinear Functionals," in Probability and Statistics: Essays in Honor of David A. Freedman, eds. D. Nolan, and T. Speed, pp. 335–421, Beachwood, OH: Institute of Mathematical Statistics. [1309,1314]
Schnabel, Z. E. (1938), "The Estimation of the Total Fish Population of a Lake," The American Mathematical Monthly, 45, 348–352. [1309]
Sekar, C. C., 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. [1311]
Stoklosa, J., and Huggins, R. M. (2012), "A Robust p-spline Approach to Closed Population Capture–Recapture Models with Time Dependence and Heterogeneity," Computational Statistics & Data Analysis, 56, 408–417. [1309]
Tilling, K., and Sterne, J. A. (1999), "Capture-Recapture Models Including Covariate Effects," American Journal of Epidemiology, 149, 392–400. [1309,1310,1311,1313,1317]
Tsatis, A. A. (2006), Semiparametric Theory and Missing Data, New York: Springer. [1312,1313]
Tsbyakov, A. B. (2008), Introduction to Nonparametric Estimation, New York: Springer. [1314]
van der Laan, M. J., and Robins, J. M. (2003), Unified Methods for Censored Longitudinal Data and Causality, New York: Springer. [1309,1312,1313]
van der Laan, M. J., and Rose, S. (2011), Targeted Learning: Causal Inference for Observational and Experimental Data, New York: Springer. [1316]
vanderLaan,M.J.,andRubin,D.(2006),"TargetedMaximumLikelihoodLearning,"TheInternationalJournalofBiostatistics,2,1–40.[1316]
vanderLaan,M.J.(2009),"PartIII:SemiparametricStatistics,"inLectures onProbabilityTheoryandStatistics,eds.P.Bernard,pp.331–457,Berlin:Springer.[1312,1313]
——— (2014), "Higher Order Tangent Spaces and Influence Functions," Statistical Science, 29, 679–686. [1309]
van der Vaart, A. W. (2002b), "Semiparametric Statistics," in Lecture Notes in Mathematics (Vol. 1781), Berlin: Springer. [1312,1313]
Vansteelandt, S., VanderWeele, T. J., Tchetgen, E. J., and Robins, J. M. (2008), “Multiply Robust Inference for Statistical Interactions,” Journal of the American Statistical Association, 103, 1693–1704. [1314]

Vardi, Y. (1985), “Empirical Distributions in Selection Bias Models,” The Annals of Statistics, 13, 178–203. [1310]

Westreich, D., and Cole, S. R. (2010), “Invited Commentary: Positivity in Practice,” American Journal of Epidemiology, 171, 674–677. [1311]

Wu, Y., and Yang, P. (2019), “Chebyshev Polynomials, Moment Matching, and Optimal Estimation of the Unseen,” The Annals of Statistics, 47, 857–883. [1313]

Yee, T. W., Stoklosa, J., and Huggins, R. M. (2015), “The VGAM Package for Capture-Recapture Data using the Conditional Likelihood,” Journal of Statistical Software, 65, 1–33. [1309]

Yip, P. S., Wan, E. C., and Chan, K. S. (2001), “A Unified Approach for Estimating Population Size in Capture-Recapture Studies with Arbitrary Removals,” Journal of Agricultural, Biological, and Environmental Statistics, 6, 183–194. [1309]

You, Y., van der Laan, M., Collender, P., Cheng, Q., Hubbard, A., Jewell, N. P., Hu, Z. T., Mejia, R., and Remais, J. (2021), “Estimation of Population Size based on Capture Recapture Designs and Evaluation of the Estimation Reliability,” arXiv preprint arXiv:2105.05373. [1311]

Zheng, W., and van der Laan, M. J. (2010), “Asymptotic Theory for Cross-Validated Targeted Maximum Likelihood Estimation.” U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 273. [1314]

Zwane, E., and van der Heijden, P. (2005), “Population Estimation using the Multiple System Estimator in the Presence of Continuous Covariates,” Statistical Modelling, 5, 39–52. [1309]