Capture-Recapture with Covariates: The Bayesian Logistic Capture-Recapture Model with Extensions

Robert Edward Granger

Submitted to the faculty of the Univesity Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Statistics,
Indiana University
May 2024 (set this to month in which all requirements are fulfilled in title.sty)

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee:

Daniel Manrique-Vallier, Ph.D.

Julia Fukuyama, Ph.D.

Amanda Mejia, Ph.D.

Roni Khardon, Ph.D.

May 1, 2024 (change this to defense date, it's in title.sty)

Robert Edward Granger

Capture-Recapture with Covariates: The Bayesian Logistic Capture-Recapture Model with Extensions

Capture-recapture refers to a series of methods that are used to estimate the size of a population from at least two incomplete, matched lists. With human populations, heterogeneity in the capture probability is a serious concern as it can impact the estimation process. This proposal outlines a framework for incorporating heterogeneity through known covariates into the estimation process and develops an extensible model under this framework which we call the Bayesian Logistic Regression model for Capture-Recapture (BLRCR). The posterior distribution is estimated using the complete likelihood through a Markov Chain Monte Carlo (MCMC) algorithm. Through a careful examination of the problem and a series of demonstrations, three important areas of future research become apparent for the dissertation project. First, we should consider the impact of the specification of the covariate distribution and explore additional non-parametric solutions. Second, we should consider the unobserved heterogeneity and specifically how to implement a procedure that accounts for this violation in the conditional independence assumption. Third, we should consider the impact of variable selection.

Contents

1 Introduction

Capture-Recapture (CR) refers to a series of methods that are used to estimate the size of a population from at least two incomplete, matched lists. While one could simply count the number of unique names between the lists, this would provide only a lower bound estimate of the population size. Presumably, there are individuals who were not captured in any of the lists. CR methods use assumptions along with patterns in the data to create estimates for how many individuals were missed and therefore inherently the size of the total population.

Most early applications of capture-recapture, also called mark and recapture, were in the context of animal populations where a number of animals were initially captured, marked or tagged, and then released. This was followed up with another capturing occasion where animals were captured and checked for a marking. A CR dataset is then a listing of each unique animal's capture pattern after all sampling occasions have taken place. CR methods have been used in a wide array of additional subject areas including ecology (Henderson and Southwood, 2016), epidemiology (Goldberg and Wittes, 1978; Baker, 1990), U.S. Census adjustments (Darroch et al., 1993), and estimating the number of pages on the internet (Lawrence and Giles, 1998; Khabsa and Giles, 2014). Of particular interest to this dissertation is the application of this methodology to human populations, which are sometimes referred to as multiple systems estimation (MSE). Lists of subjects may be collected by various institutions including hospitals, government agencies, or other non-governmental organizations but are left incomplete due to lack of resources, mistakes/omissions, or just the inability to obtain such information. These lists can then be combined and constructed to form a dataset for CR. For example, Zwane and van der Heijden (2005) attempts to determine the number of Dutch children born with a neural tube defect (NTD) in the year 2000 by analyzing three national databases tracking these occurrences. For multiple reasons including risk assessment of the pregnancy and

omissions, not all children with NTDs are reported to each database. Another example is the counting of casualties from various conflicts including Colombia (Guberek et al., 2010; Manrique-Vallier, 2016), Guatemala (Ball et al., 2000; Ball and Price, 2018), Kosovo (Ball et al., 2002; Manrique-Vallier, 2016), Peru (Ball et al., 2003), and many others. Casualty counts by a reporting agency may only be partially collected because of non-cooperation from victims due to lack of trust, danger posed to data collectors, and/or destroyed infrastructure (Manrique-Vallier et al., 2020).

CR methods often come with many assumptions including a closed population, independence between lists, and homogeneity in the capture probability between individuals. In particular, the aim of the dissertation is to extend methods that target the homogeneity assumption, the requirement that all individuals have the same probability of capture on each list regardless of any personal attributes. While this may be a palatable assumption for some animal populations, this is unlikely with human populations. Different characteristics like age or social status may influence the probability of their capture. If data is present that can fully explain the heterogeneity, one could incorporate this by stratifying before applying their CR method of choice (Sekar and Deming, 1949; Manrique-Vallier et al., 2019). This is only possible if the covariate is discrete or can be made discrete. It can also lead to issues of sparsity. An alternative is to include the covariates in a model like regression (Alho, 1990; Baker, 1990; Bonner and Schofield, 2014; King et al., 2016). Regardless, this only addresses the observable heterogeneity and ignores heterogeneity that may exist but covariate information is missing or is not fully adequate. Some approaches have been proposed to account for unobservable heterogeneity by adapting the Rasch model (Darroch et al., 1993) or latent class memberships (Manrique-Vallier, 2016).

This dissertation develops a fully Bayesian procedure for the CR problem that has a number of desirable characteristics:

1) allows for multiple lists;

2) is resistant to sparsity;

3) can be informed through the use of priors;

4) uses discrete and continuous covariates to account for observable heterogeneity;

5) can also account for unobservable heterogeneity;

6) extensible;

Estimating the size of the population tends to be highly susceptible to the structure of the modeling when dealing with CR methods. Therefore, implementing covariates in an improper way or not adequately accounting for the unobservable heterogeneity can bias the inference. As a result, we propose developing a method that is fully Bayesian approach and uses the full likelihood as suggested by Pollock (2002). We begin by developing a framework for CR that incorporates covariates, a framework that for all intents and purposes is identical to the complete likelihood specification in (King et al., 2016). Using this framework, we specify a model that uses conditionally independent logistic regressions which we term the Bayesian Logistic Regression Capture-Recapture (BLRCR) model.

While the BLRCR model incorporates individual covariates and thus accounts for observable heterogeneity, it has a few shortcomings which are addressed in subsequent sections.

1) Since covariates are missing at least for the unobserved individuals, the use of a full likelihood approach requires the specification of a distribution for the covariates. As we will see in Section 3.3, misspecifying the distribution can lead to incorrect inference. Like previous approaches, we apply a normal distribution; however, we propose a new method of using non-parametric approaches to the covariates such as an infinite mixture of normal distributions.

2) Unobservable heterogeneity can bias the inference, and hence we need a way to account for it. From our perspective, we view the idea of unobservable heterogeneity as an issue of omitted covariates. We borrow the idea of individuals belonging to latent classes impacts the probability of capture(Manrique-Vallier, 2016), but implement it in the regression procedure through an additional intercept term.

3) Determining which covariates to include and how each covariate should be included in the regression will ultimately impact the inference. There are multiple ways to account for this issue, but we explore it through the lens of variable selection.

The dissertation proceeds as follows: Section 2 partially reviews the relevant capture-recapture literature. Section 3 details the BLRCR model along with the three extensions described above. This includes the derivation of the posterior and a Markov Chain Monte Carlo (MCMC) estimation procedure. Section 4 contains an analysis of the procedure using a series of simulations. Section 5 looks at Example 1. Section 6 looks at Example 2. Section 7 concludes with a summary of the dissertation and future directions of work in this area.

2 Literature Review

Capture-Recapture (CR) refers to a series of methods that are used to estimate the size of a population from at least two incomplete, matched lists. This naming comes from the process of first capturing/marking a number of organisms (capture) and following up with a second capturing occasion (recapture). One then estimates the population size, N, of the organism by comparing the number of organisms marked in each sample, n_1 and n_2, and the organisms that were matched in both samples, n_{12}. Intuitively, the proportion of the total population that is initially captured/marked, n_1/N, should be equal to the proportion of captured/marked individuals that are recaptured, n_{12}/n_2. Of course, this only makes sense if we believe the lists are independent. Rearranging and solving for N leads to the formula:

$$\hat{N} = \frac{n_1 \cdot n_2}{n_{12}}.$$

Although this estimator was used as early as 1783 by the mathematician Pierre-Simon Laplace for the purpose of estimating the French population (Schaefer, 1951), this estimator is commonly referred to as the Lincoln-Petersen estimator after its early use by Frederick C. Lincoln on migratory waterfowl (Lincoln, 1930) and C.G. Johannes Petersen on the fish species plaice (Petersen, 1895). The Lincoln-Petersen estimator is relatively simple to compute, but is limited to only two lists with the assumption of a closed population. A closed population means the population does not change throughout the sampling occasions. This estimator also makes two other intricately related assumptions:

1) independence between lists,

2) homogeneity in capture probabilities between individuals.

The first assumption of independence between lists means the probability an individual is captured on one list does not dependent on the probability that individual is captured on

another list. The second assumption means the probability of any two individuals being captured on a list is the same. When these assumptions are violated, it can lead to bias in our population estimates. Addressing violations of these assumptions has a long history and continues to be looked into today.

2.1 Early Approaches to List Dependency and Homogeneity

Capture-recapture solutions with multiple lists appear at least as early as the 1920s (Geiger and Werner, 1924; Schnabel, 1938; Darroch, 1958). All use a model for estimating the size of a population from multiple lists with list independence. If this assumption were to hold, then the probability an individual appears on one list would not impact the probability that the individual appears on another list. Perhaps for carefully controlled experiments, list independence can be assumed, but for most real applications, this is unlikely. For example with human rights data, this may be violated as psychological research indicates survivors of human rights abuses can benefit in their mental health by sharing their testimony. These individuals may in turn be encouraged to share their experiences to multiple outlets. Furthermore, there is evidence with the Guatemalan Civil War that several popular movements groups encouraged their social bases to provide testimony of human rights abuses to all of three of the data collecting organizations (Ball et al., 2000). These conditions lead to positive list dependence and therefore an underestimate of the population size. On the other hand, there is also the potential that certain groups of individuals may feel uneasy sharing their testimony with non-aligned data collection groups, leading to negative dependence and an overestimate of the population size. For example, religious members of the Catholic faith were more apt to give their testimony to the Catholic researchers than they were to the political left, non-governmental organizations (Manrique-Vallier et al., 2020).

In the early 1970s, two approaches were developed to address list dependency. Fienberg

(1972) addressed the issue of list dependency directly by modeling interactions between lists using hierarchical log linear models under a multinomial sampling scheme. On the other hand, Sanathanan (1972) also used a multinomial sampling scheme but instead of modeling list dependence directly, she modeled it as unobserved heterogeneity in the individuals under a conditional likelihood approach. As pointed out in several places (Darroch et al., 1993; Fienberg et al., 1999; Manrique-Vallier et al., 2020), list dependency and heterogeneity between individuals are intricately related such that list dependency can be viewed as a result of heterogeneity between individuals. This leads to the second assumption, homogeneity, which refers to all individuals having the same capture probability within a list. For a variety of reasons, we may expect certain individuals to be more likely to be captured because of characteristics like geographic area, age, or their status within the community.

This heterogeneity among individuals may be unobserved, observed through covariates, or a combination. In the 1990s, various methods were proposed to account for unobserved heterogeneity using adaptations to log-linear and mixture models (Darroch et al., 1993; Agresti, 1994; Pledger, 2000). Later, Bayesian methods based on latent class models were presented that allowed for a more flexible approach to heterogeneity (Fienberg et al., 1999; Manrique-Vallier and Fienberg, 2008; Manrique-Vallier, 2016). These methods seek out patterns in the capture histories in order to infer for unobserved heterogeneity. When the researcher has accompanying discrete covariate data, a suggestion noted by Darroch et al. (1993), is to first stratify the data based on the observable heterogeneity (Sekar and Deming, 1949), and then perform the desired method within that stratification. An example of this this implementation can be found in Manrique-Vallier et al. (2019), where the authors estimate the number of casualties in the internal conflict in Peru (1980-2000) based on geographic stratifications. Stratification becomes problematic in the presence of continuous covariates. When continuous covariates are present, the researcher must make difficult decisions on how to discretize the covariates which leads to assigning potentially

arbitrary cutoffs. If your stratification scheme results in too many strata, the number of observed individuals within each strata may be small, resulting in loss of inferential power and identifiability. These issues can arise even without continuous covariates if there are many discrete variables with many unique levels.

2.2 Capture-Recapture with Covariates

An alternative to stratification is to use the covariates within the modeling process like with regression. Pollock et al. (1984) suggested the use of logistic regression with environmental and individual level covariates. Huggins (1989); Alho (1990) independently proposed a similar but slightly different approach that uses conditional maximum likelihood logistic regression through a two step process. First, the coefficients for the logistic regression model are obtained by regressing whether the individual was captured against the covariates under the condition the individual is observed at least once. If we perform the regression without taking into account that only observed individuals appear in the dataset, i.e., some individuals are missing, the coefficients become biased. After obtaining the coefficients, we fit the probability that each individual is captured on each list, and the probability an individual is missing, θ_i, is computed. The second step uses the estimated probability of missing by plugging it into the Horvitz and Thompson (1952) estimator,

$$\hat{N} = \sum_{i}^{n} \frac{1}{\theta_i},$$

to obtain an estimate for N. Intuitively, the estimator inverts the probability of missing to account for the number of similar individuals that would have been unobserved. This intutition implies that the estimator will not perform well when the covariates for the observed individuals vary significantly from the unobserved individuals. Nevertheless, Alho (1990) shows this estimator to be consistent with large samples and develops an asymptotic approach to estimating the variance.

In these models, the probability an individual is captured on each list is conditionally independent of the other lists conditional on the covariates. While covariates may explain some of the heterogeneity, it is unlikely that all heterogeneity between individuals can be explained in such a fashion and that some list dependency would not remain. Zwane and van der Heijden (2005) use multinomial logistic regression with a design matrix intended to directly model the dependence between the lists. One could theoretically incorporate additional rows into the design matrix to account for unobserved heterogeneity as in Darroch et al. (1993).

Since the covariates will be missing for any of the unobserved individuals, any approach that uses the complete likelihood would require the covariate distribution to be specified. All of the previous mentioned methods sidestep this need by maximizing the conditional likelihood instead of the full likelihood. Stoklosa et al. (2011) takes this a step further by comparing a partial likelihood approach, i.e, the number of recaptures after the first capture, with the conditional approach. They find a loss of efficiency but argue that it allows more flexibility in modeling. Later, Yee et al. (2015) presented a simple way to implement conditional likelihood methods using the `VGAM` package in the programming software `R`. There is some early usage of a full likelihood approach appearing in work that combines line transact theory with capture-recapture for population estimation (Alpizar-Jara and Pollock, 1996; Borchers et al., 1998). Since the full likelihood approach requires specifying a distribution for the covariates, this presents the difficulty of potentially estimating additional parameters regarding the covariate distribution and/or integrating out the covariates to obtain an estimate for N. While more challenging, Pollock (2002) speculates the full likelihood approach could result in better precision and suggests a possible solution may lie with Bayesian methods.

2.3 Bayesian Methods for Capture-Recapture

Perhaps the earliest paper to utilize Bayesian methods is Roberts (1967) in the evaluation of "stopping rules" with exactly two samples. The "stopping rule" refers to how the sample of individuals from the population will be collected, and in particular, what criterion must be met before concluding the sampling occasion. For example, a researcher may choose a fixed sample size and will stop sampling subjects when this sample size is reached. On the other hand, the researcher may attempt to sample every subject within a fixed time period. Depending on the chosen stopping rule, the likelihoods can differ and thus the estimator. See Chapman (1954); Darroch (1958) for a non-Bayesian look at the problem. These different stopping rules were once again considered in the Bayesian approaches of Castledine (1981) and extended to include more than two lists. Further, the authors assume the probability of capture is constant between individuals but allowed to vary across time (see model M_t in Otis et al. (1978)). The impact of various prior specifications is explored including the use of the Beta prior for the probability of capture and the Jeffreys prior (Jeffreys, 1967) for N. Under similar modelling, estimates of the posterior distribution were obtained using empirical Bayes and Bayes empirical Bayes (Smith, 1991), and Markov Chain Monte Carlo (MCMC), Gibbs sampling, (George, 1992) methodology.

Bayesian approaches to unobserved heterogeneity were first introduced to the multiple-recapture problem in Fienberg et al. (1999). The authors use a fully Bayesian hierarchical approach through the use the Rasch model and develop a MCMC procedure for obtaining samples from the posterior. Manrique-Vallier and Fienberg (2008) also employ a fully Bayesian hierarchical approach but use a Grade of Membership model (Woodbury et al., 1978) to account for the unobserved heterogeneity. Individuals are soft or partially clustered into classes, i.e. mixed membership, with each class having an estimated probability of capture per list. A simplification to the partial membership approach is to use the Latent Class Model (LCM) where individuals are hard clustered into

groups. (Manrique-Vallier, 2016) uses a Bayesian Nonparametric Latent Class Model (NPLCM) by using a potentially infinite number of hidden classes through a stick-breaking prior (Dunson and Park, 2008). Posterior samples are obtained via a Gibbs sampler.

Naturally, when deriving a posterior distribution, one would use the full or "complete" likelihood. This has the benefit of 1) being the "correct" way to compute the posterior based on a specified generative process and 2) allows us to extend our model without fear of compounding any errors from modifications/approximations. Unfortunately, the use of a full likelihood approach in the presence of covariates poses a particularly challenging problem as a distribution for the covariates must be specified and the posterior often becomes intractable. The early approach of using the conditional likelihood instead of the full likelihood avoids this problem but recently several other solutions have been presented. These include the use of a reversible jump MCMC (King and Brooks, 2008), data augmentation (Royle et al., 2007; Royle, 2009), and replacing the full likelihood with a "semi-complete" data likelihood approach (King et al., 2016). The semi-complete likelihood uses the complete likelihood for the observed individuals and a marginal likelihood for the unobserved individuals. Both approaches still require estimating the probability of missing, which is a difficult integral. While (King et al., 2016) evaluate the integral using Gauss-Hermite quadrature, Bonner and Schofield (2014) estimates the probability of missing with a Monte Carlo within MCMC step. The efficiency of this approach depends on the size of the Monte Carlo simulated covariate distribution with a larger sample giving a better approximation, but at the cost of reducing the efficiency of the algorithm. It should be noted that whether we use a full likelihood or even a semi-complete likelihood, a covariate distribution must be specified. The examples provided often assume a normal distribution, but Royle (2009) does consider alternative distributions. These alternative distributions do reveal some sensitivity to the choice of covariate distribution, which will be explored further in subsection 3.3.

3 Capture-Recapture with Covariates: The Bayesian Logistic Capture-Recapture Model with Extensions

In this section, we develop a Bayesian procedure for the multiple-list capture-recapture (CR) problem with covariates. Perhaps the most ubiquitous technique in solving the CR problem with multiple-lists is through the frequentist technique of log-linear models (Fienberg, 1972). Using multiple lists ($J \geq 2$) can be beneficial in that we have more detailed patterns, which can lead to better inference. The downside is that as the number of lists increases, the number of possible patterns grows exponentially. For example, with just two lists, there are only four possible patterns: 1) the individual shows up on both lists, 2) the individual shows up only on list 1, 3) the individual shows up only on list 2, or 4) the individual shows up on neither list. The number of possible patterns can be calculated as 2^J, so a dataset with 30 lists would have over a billion possible patterns. Not only could this become computationally expensive, but it also leads to issues of sparsity, i.e., many of the potential patterns will not appear in the data. In Manrique-Vallier (2016), they analyze a dataset about killings in Casanare, Colombia which contains 15 lists. They note that only 70 of the potential $2^{15} = 32,768$ capture patterns are present and as a result were unable to successfully compute a solution using log-linear models.

Even data with much fewer lists can still have issues of sparsity. To get around this issue, we present the Bayesian Logistic Regression Capture-Recapture model that relies on an assumption of conditional independence. First, since the method is Bayesian, we have the added benefit of allowing the practitioner to insert prior belief or knowledge into the estimation procedure. Bayesian methods also are known to assist with sparsity in that they essentially "create" data to fill in the sparse areas. In addition, the assumption of conditional independence, while strong, allows for a reduction in the complexity and sparsity issue. The conditional independence assumption states that given the covariates, the probability of capture on one list is unaffected by another list.

Unlike most popular existing methods, the method presented here introduces the ability to insert covariates to guide in the estimation of the capture probability. Instead of relying on stratification that essentially only allows for discrete covariates, we implement conditionally independent logistic regressions on each of the lists to determine the probability of capture (or inversely, the probability of non-capture). Typically with regression, the distribution of the covariate is irrelevant so it does not matter whether it is discrete or continuous. Unfortunately, since we're dealing with what amounts to a missing data problem, but the missing data is not missing at random (Rubin, 1976), the distribution of the covariate matters (see subsection 3.3). Nevertheless, while we must be careful how we model the distribution of the covariate, any type of covariate, continuous or discrete, can be used.

Finally, a key concern of any capture-recapture framework is that the estimation is highly dependent on how the model is structured. Different situations may present important characteristics of the problem that need to be carefully considered by the practioner. With that in mind, the desire is to create a methodology that is extensible. Through the use of the full likelihood and a data augmentation approach, we are able to add extensions more readily including more robust ways of handling heterogeneity. In particular, we may be concerned with both "observable" heterogeneity, i.e. the heterogeneity that can be detected through the use of covariates, but we may also be concerned with "unobservable" heterogeneity, i.e, the heterogeneity that persists but for which no covariate information exists. One such method, presented in subsection 3.4, allows for the modeling of this unobservable heterogeneity through the use of latent classes. This assumption also happens to break the perhaps unpalatable assumption of conditional independence mentioned before. Similarly, the covariates that are present may be numerous or may affect the estimation procedure in non-linear ways, so we extend the model to include a variable selection procedure (see subsection 3.5). Additional extensions may be warranted given the problem, but the estimation procedure that uses data augmentation allows for extensions to be added relatively easily.

To summarize, we begin with a Bayesian framework for the capture-recapture problem with covariates (see subsection 3.1). Next, we propose a specific model following this procedure which we call the Bayesian Logistic Regression Capture-Recapture (BLRCR) model and estimate the posterior through a Markov Chain Monte Carlo (MCMC) algorithm (see subsection 3.2). The three remaining subsections discuss extensions to specifying the distribution of the covariates (subsection 3.3), discovering unobserved heterogeneity (subsection 3.4), and variable selection (subsection 3.5).

3.1 Background and Preliminaries

3.1.1 The Classical Capture-Recapture Approach

We begin with the multinomial multiple-recapture framework first presented in Darroch (1958) and utilized by numerous past and more recent works (Sandland and Cormack, 1984). This section summarizes this framework, heavily relying on the notation of Manrique-Vallier (2016). The objective of this capture-recapture framework is to estimate the unknown size, N, of a population of individuals assuming the population size remains unchanged throughout the capturing occasions, i.e. a closed population, and that the captured individuals can be matched perfectly. Individuals are captured (sampled) through the use of $J \geq 2$ lists. If individual, i, is captured on list j, then $y_{ij} = 1$ with $y_{ij} = 0$ otherwise. When aggregated across lists, we refer to these values as capture vectors, $\boldsymbol{y_i} = (y_{i1}, ..., y_{iJ})$. If an individual is not identified on any of the J lists, $\boldsymbol{y_i} = \mathbf{0} \equiv (0, ..., 0)$, then that individual is considered "unobserved" or "missing". The number of unobserved individuals, n_0, plus the number of individuals that are captured on at least one list, n, is equal to the size of the population, i.e, $N = n_0 + n$.

Each individual's capture vector, $\boldsymbol{y_i}$, is generated from probability distribution, $f(\boldsymbol{y}|\theta)$ for all $i = 1, ..., N$. Reordering the individuals such that the observed individuals are $1, ..., n$ and the unobserved individuals are $n + 1, ..., N$ leads to the following joint likelihood

$$p(\mathcal{Y}_{obs}|N,\boldsymbol{\theta}) = \binom{N}{n} f(\mathbf{0}|\boldsymbol{\theta})^{N-n} \prod_{i=1}^{n} f(\boldsymbol{y_i}|\boldsymbol{\theta}) I(N \geq n), \tag{1}$$

where $\mathcal{Y}_{obs} = (\boldsymbol{y}_1, ..., \boldsymbol{y_n})$, the observed capture vectors. For the classical CR problem, the only data is the observed capture histories, $\mathcal{Y}_{obs}$. The parameters of interest are N and $\boldsymbol{\theta}$. We place a prior distribution, $p(N, \boldsymbol{\theta})$, on the parameters with the objective of computing the posterior distribution,

$$p(N,\boldsymbol{\theta}|\mathcal{Y}_{obs}) \propto p(\mathcal{Y}_{obs}|N,\boldsymbol{\theta}) p(N,\boldsymbol{\theta}). \tag{2}$$

3.1.2 Capture-Recapture with Covariates

We expand the framework from the previous section by allowing each individual's capture probability to be dependent on a matrix of covariates, $\mathcal{X}$. Let x_{ih} represent the value of covariate $h \in 1, ..., H$ for individual $i \in 1, .., N$. Since some individuals are not captured, the covariate information for these individuals is lost. Let $\mathcal{X}_{obs} = (\boldsymbol{x_i}, ..., \boldsymbol{x_n})$ be the covariate data on the $i, ..., n$ individuals that are observed and let $\mathcal{X}_{mis} = (\boldsymbol{x_{n+1}}, ..., \boldsymbol{x_N})$ be the covariate data on the $n+1, ..., N$ individuals that are not observed.

We update the joint likelihood in Equation 1 to include these covariates,

$$\begin{aligned} p(\mathcal{Y}_{obs}, \mathcal{X}_{obs}|N,\theta,\boldsymbol{\phi},\mathcal{X}_{mis}) &= p(\mathcal{Y}_{obs}|N,\theta,\boldsymbol{\phi},\mathcal{X}_{obs},\mathcal{X}_{mis}) p(\mathcal{X}_{obs}|\boldsymbol{\phi}) \\ &\propto \binom{N}{n} \prod_{i=1}^{n} f(\boldsymbol{y_i}|\boldsymbol{\theta},\mathcal{X}_{obs}) \prod_{n+1}^{N} f(\mathbf{0}|\boldsymbol{\theta},\mathcal{X}_{mis}) I(N \geq n) \cdot g(\mathcal{X}_{obs}|\boldsymbol{\phi}). \end{aligned} \tag{3}$$

Once again, in order to complete the Bayesian model, priors must be assigned to the unknown parameters $\boldsymbol{\theta}$ and N; however, we also must consider the distribution of the observed and missing covariates along with the parameter(s) governing their distribution,

$\boldsymbol{\phi}$. The joint posterior can be written as,

$$\begin{aligned} p(N, \boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{X}_{mis} | \mathcal{Y}_{obs}, \mathcal{X}_{obs}) &\propto p(\mathcal{Y}_{obs}, \mathcal{X}_{obs} | N, \boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{X}_{mis}) p(N, \boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{X}_{mis}) \\ &= p(\mathcal{Y}_{obs}, \mathcal{X}_{obs} | N, \boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{X}_{mis}) p(\mathcal{X}_{mis} | N, \boldsymbol{\phi}) p(N, \boldsymbol{\theta}, \boldsymbol{\phi}) \end{aligned} \tag{4}$$

Unfortunately, Equation 4 does not allow us to simply compute the posterior by conditioning on the covariates as in classical regression. Let $\mathcal{Y} = [\mathcal{Y}_{obs}, \mathcal{Y}_{mis}]_{N \times J}$ and $\mathcal{X} = [\mathcal{X}_{obs}, \mathcal{X}_{mis}]_{N \times H}$ represent the complete data; however, the unobserved portions of each of these matrices are fundamentally different. Conceptually, $\mathcal{Y} = [\mathcal{Y}_{obs}, \mathcal{Y}_{mis}]$ is decomposed into observed and unobserved components; however, $\boldsymbol{y_i} = \mathbf{0}$ for all $\boldsymbol{y_i} \in \mathcal{Y}_{mis}$,i.e., the values of each row are known. On the other hand, we do not know the values of the missing covariates, $\mathcal{X}_{mis}$. This complicates our ability to compute the posterior, which could have been a simple case of conditioning on the covariates as in classical regression. Because of the missing covariates, we specify a distribution for all covariates,

$$\boldsymbol{x_i} \overset{iid}{\sim} \boldsymbol{g}(\boldsymbol{\phi}). \tag{5}$$

To further complicate matters, the observed and unobserved covariates almost surely do not follow the same distribution as they are not missing at random. While $\boldsymbol{x_i} \in \mathcal{X}_{obs}$ or $\boldsymbol{x_i} \in \mathcal{X}_{mis}$ is defined through the observational status of its corresponding $\boldsymbol{y_i}$, all $\boldsymbol{x_i}$ in each set is drawn independently from Equation 5. Nevertheless, it is not the case that $\boldsymbol{x_i}|\boldsymbol{y_i} \neq \mathbf{0}$ or $\boldsymbol{x_i}|\boldsymbol{y_i} = \mathbf{0}$ will necessarily have this same distribution. In other words, if we desire samples of our missing covariates, it would be incorrect to sample simply from $g(\boldsymbol{\phi})$, but instead these values must sampled conditioned on the individual being missing.

3.2 The Bayesian Logistic Regression Capture-Recapture Model

3.2.1 The BLRCR Model

Using the framework from subsection 3.1.2, we implement the Bayesian Logistic Regression Capture-Recapture (BLRCR) model which uses independent logistic regressions to estimate the capture probabilities on each list. Suppose y_{ij} and x_{ih} are generated in the following way:

$$y_{ij}|\boldsymbol{x_i} \overset{ind}{\sim} \text{Bernoulli}(\lambda_{ij}(\boldsymbol{x_i})) \quad \text{for } i = 1, ..., N \text{ and } j = 1, ..., J \tag{6}$$

$$\boldsymbol{x_i} \overset{iid}{\sim} \boldsymbol{g}(\boldsymbol{\phi}) \quad \text{for } i = 1, ..., N, \tag{7}$$

where

$$\lambda_{ij}(\boldsymbol{x_i}) = \sigma(\boldsymbol{x_i}^T\boldsymbol{\beta_j}) = \frac{1}{1 + e^{-\boldsymbol{x_i}^T\beta_j}}, \tag{8}$$

and $\boldsymbol{g}(\boldsymbol{\phi})$ is the distribution of the H covariates. The capture probability that individual i appears on list j is equal to λ_{ij}, and this value can be calculated with a nonstochastic transformation of $\boldsymbol{x_i}$ and $\boldsymbol{\beta_j}$ through the sigmoid function (see Equation 8). In the linear term, $\boldsymbol{x_i}^T\boldsymbol{\beta_j}$, a different set of $\boldsymbol{\beta_j}$ covariates are used to determine each λ_{ij} implying a total of $k \times (h + 1)$ coefficients with the inclusion of an intercept. Notice this setup implies each individual's capture pattern is independent conditional on the covariates, i.e, $p(\boldsymbol{y_i}|\boldsymbol{x_i}) = \prod_{i=j}^{J} p(y_{ij}|x_i, \theta)$.

King et al. (2016) proposes a similar model in their first example titled "Continuous individual covariates." An important distinction though is the allowance, for each covariate, to have a different set of coefficients, $\boldsymbol{\beta_j}$. King et al. (2016) justifies a single coefficient per covariate as the model they develop is in the context of animal populations. In animal

populations, each list is a different capturing from the same population but at various time points. In order to assume a closed population, the lists should be collected in a relatively short time period. As a result, the author argues that any time varying effect from the covariates should be limited. On the other hand, we approach this problem in the context of human populations, where lists are not necessarily different time points but instead different data collectors or databases. As a result, we expect individual characteristics (covariates) to have different impacts on the capture probability depending on the list.

In order to complete the Bayesian model, priors must be assigned to unknown parameters N, $\boldsymbol{\beta_j}$, and $\boldsymbol{\phi_h}$. For N, we use the Jeffrey's prior (Jeffreys, 1967), $p(N) = \frac{1}{N}$, which conveniently results in a negative binomial distribution for the conditional posterior distribution of N. Other choices of priors typically result in more complicated estimation, especially with data augmentation (see King et al. (2016) for a discussion). For the $\boldsymbol{\beta_j}$ coefficients, we assign a multivariate normal prior to the set of coefficients for each list with mean of $\boldsymbol{b} \in \mathcal{R}^{H+1}$ and covariance of $\boldsymbol{B} \in \mathcal{R}^{(H+1)\times(H+1)}$. Other choices of prior can be used as the problem reduces to Bayesian logistic regression after augmenting the missing data. We select the multivariate normal prior as it is the same prior used in the Bayesian logistic regression Monte Carlo Markov Chain (MCMC) sampling scheme proposed in (Polson et al., 2013), which makes implementation simple. We can be flexible in our choice of prior distribution for the $\boldsymbol{\beta_j}$ coefficients as long as a suitable method exists for drawing samples from the conditional posterior distribution for $\boldsymbol{\beta_j}$. Lastly, for now, assume $\boldsymbol{\phi_h}$ is known and thus can be treated as a hyperparameter (this will be further addressed in subsection 3.3).

Plugging in the likelihood distribution as described in Equation 6 and the aforementioned

priors into Equation 4 yields the posterior

$$
\begin{aligned}
p(N, \boldsymbol{\beta}, \mathcal{X}_{mis} | \mathcal{Y}_{obs}, \mathcal{X}_{obs}) \propto & \left[\binom{N}{n} \prod_{i=1}^{n} \prod_{j=1}^{J} \lambda_{ij}(\boldsymbol{x_i})^{y_{ij}} (1 - \lambda_{ij}(\boldsymbol{x_i}))^{1-y_{ij}} \prod_{i=n+1}^{N} \prod_{j=1}^{J} (1 - \lambda_{ij}(\boldsymbol{x_i})) \right] \\
& \times \left[\prod_{i=1}^{n} \boldsymbol{g}(\boldsymbol{x_i} | \boldsymbol{\phi_h}) \right] \times \left[\prod_{i=n+1}^{N} \boldsymbol{g}(\boldsymbol{x_i} | \boldsymbol{\phi_h}) \right] \times \left[\frac{1}{N} \right] \\
& \times \left[\prod_{j=1}^{J} \left(\frac{1}{2\pi} \right)^{H/2} |\boldsymbol{B}|^{-1/2} e^{-\frac{1}{2}(\boldsymbol{b} - \boldsymbol{\beta_j})^T \boldsymbol{B}^{-1} (\boldsymbol{b} - \boldsymbol{\beta_j})} \right]. \qquad (9)
\end{aligned}
$$

3.2.2 Estimation of the BLRCR

An exact analytic solution for the posterior is intractable, so we implement the MCMC algorithm of Gibbs Sampling, where sequential draws from $\boldsymbol{\beta}$, N, and $\mathcal{X}_{mis}$ are taken conditional on all other parameters.

1) Sample $\boldsymbol{\beta}$. For this stage $\mathcal{X}_{mis}$ and $\mathcal{X}_{obs}$ are both known implying N is known as well. Therefore the sampling equation reduces to

$$
p(\boldsymbol{\beta} | \mathcal{Y}, N, \mathcal{X}_{mis}, \mathcal{X}_{obs}) = p(\boldsymbol{\beta} | \mathcal{Y}, \mathcal{X}), \qquad (10)
$$

which is simply the posterior distribution of Bayesian logistic regression. Posterior samples can be obtained from Equation 10 by first sampling a latent variable from the Polya-Gamma distribution and using this latent variable in the mean and covariance function of a multivariate normal (Polson et al., 2013).

2) Sample N and $\mathcal{X}_{mis}$. Adding up the number of missing covariates, n_0, with the number of observed covariates, n, fully determines $N = n_0 + n$. This issue makes it impossible to compute the standard Gibbs sampling equation for N. To get around this complication, we sample the parameters simultaneously (Basu, 2001),

$$
p(N, \mathcal{X}_{mis} | \mathcal{Y}, \boldsymbol{\beta}, \mathcal{X}_{obs}) \propto p(N | \mathcal{Y}, \boldsymbol{\beta}, \mathcal{X}_{obs}) p(\mathcal{X}_{mis} | N, \mathcal{Y}, \boldsymbol{\beta}, \mathcal{X}_{obs}). \qquad (11)
$$

From Equation 11, observe the joint distribution of N and $\mathcal{X}_{obs}$ can be decomposed into two parts from which can be sampled. The first part of this equation is the joint distribution of N and $\mathcal{X}_{mis}$ marginalized over the missing covariates. The second part is the distribution of the missing covariates where the number of missing covariates, n_0, is known.

i. Sample $N \sim p(N|\mathcal{Y}, \boldsymbol{\beta}, \mathcal{X}_{obs})$.

$$\begin{aligned} p(N|\mathcal{Y}, \boldsymbol{\beta}, \mathcal{X}_{obs}) &= \int_{\boldsymbol{x_{n+1}}} \cdots \int_{\boldsymbol{x_N}} p(N, \mathcal{X}_{mis}|\mathcal{Y}, \boldsymbol{\beta}, \mathcal{X}_{obs}) d\boldsymbol{x_{n+1}} \cdots d\boldsymbol{x_N} \\ &\propto \int_{\boldsymbol{x_{n+1}}} \cdots \int_{\boldsymbol{x_N}} \left[\binom{N}{n} \prod_{i=n+1}^{N} \prod_{j=1}^{J} (1-\lambda_{ij}) \right] \left[\prod_{i=n+1}^{N} \boldsymbol{g}(\boldsymbol{x_i}|\boldsymbol{\phi}) \right] \left[\frac{1}{N} \right] d\boldsymbol{x_{n+1}} \cdots d\boldsymbol{x_N} \\ &= \frac{(N-1)!}{(N-n)!n!} \left[\int_{\boldsymbol{x}} \boldsymbol{g}(\boldsymbol{x_i}|\boldsymbol{\phi}) \prod_{j=1}^{J} (1-\lambda_{ij}) d\boldsymbol{x} \right]^{N-n} \\ &\propto \binom{N-1}{n-1} \left[\underbrace{E_{\boldsymbol{g}(\boldsymbol{\theta_h})} \left[\prod_{j=1}^{J} (1-\lambda_{ij}) \right]}_{\equiv \rho} \right]^{N-n} \end{aligned} \tag{12}$$

Instead of sampling N, sample $n_0 = N - n$.

$$\begin{aligned} p(n_0|\mathcal{Y}, \boldsymbol{\beta}, \mathcal{X}_{obs}) &\propto \binom{n_0+n-1}{n-1} \rho^{n_0} \\ &\propto \binom{n_0+n-1}{n-1} \rho^{n_0} \underbrace{(1-\rho)^n}_{=constant} \\ &= \text{NegativeBinomial}(n, 1-\rho). \end{aligned} \tag{13}$$

The distribution of n_0 follows a negative binomial with parameter n for the number of "successes" and $1-p$ as the "success" rate. In this context, a "success" is defined as an observation being unobserved. The value of ρ is defined in Equation 12 and can be computed via numerical integration or estimated via a Monte Carlo within MCMC step as in Bonner and Schofield (2014). After

computing ρ and sampling n_0 through Equation 13, find $N = n_0 + n$.

ii. Sample $\mathcal{X}_{mis}$. The missing observation, $\boldsymbol{x_i}$, is drawn independently of all other covariates, so $\boldsymbol{x_i}$ does not depend on any other $\boldsymbol{x_i} \in \mathcal{X}_{obs} \cup \mathcal{X}_{mis}$. Also, by definition, if $\boldsymbol{x_i} \in \mathcal{X}_{mis}$, then $\boldsymbol{y_i} = \mathbf{0}$. Therefore, the distribution to be sampled is

$$\begin{aligned} p(\boldsymbol{x_i}|N, \mathcal{Y}, \boldsymbol{\beta}, \mathcal{X}_{obs}) &= p(\boldsymbol{x_i}|\boldsymbol{y_i} = \mathbf{0}, \boldsymbol{\beta}) \\ &\propto \boldsymbol{g}(\boldsymbol{x_i}|\boldsymbol{\phi}) \prod_{j=1}^{J} (1 - \lambda_{ij}(\boldsymbol{x_i})) \end{aligned} \tag{14}$$

To sample from Equation 14, we use rejection sampling of a truncated distribution. To do this, first draw a sample $\boldsymbol{x_i}$ from the distribution of the covariates, $\boldsymbol{g}(\boldsymbol{\phi})$. Next, accept the sample with probability of $\boldsymbol{x_i}$ being missing, i.e, $\prod_{j=1}^{J}(1 - \lambda_{ij}(\boldsymbol{x_i}))$. If the sample is not accepted, reject it, and draw another sample from $\boldsymbol{g}(\boldsymbol{\phi})$. Repeat this until we obtain $n_0 = N - n$ missing covariates. On average, this sampling techniques requires us to draw N total covariates for each sample.

3.3 Selecting a Distribution for the Covariates

When the probability that an individual is captured at least once is dependent upon the covariates, the observed covariate distribution will differ from population covariate distribution. Simply using logistic regression on the observed data would lead to biased coefficients, which in turn would lead to bias in the population estimation. The algorithm presented in subsection 3.2 alleviates this problem by concatenating samples of the missing covariates with the observed data before estimating the coefficients. Unfortunately, this presents an additional burden on the user of specifying a distribution for the missing covariates, $\boldsymbol{g}(\boldsymbol{\phi})$.

3.3.1 Specifying a Distribution for the Covariates

If the distribution is known including parameters, then we can simply use the aforementioned algorithm. If the distribution is known except for the parameters, then we could specify that distribution along with priors on the parameters. For example, Royle (2009) uses a single covariate and specifies a normal distribution with a normal distribution prior for the mean and a gamma prior for the inverse variance. One could take a multivariate version of this approach by specifying,

$$\boldsymbol{x_i} \overset{iid}{\sim} \text{MVNormal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{15}$$

with conjugate priors,

$$\boldsymbol{\Sigma} \sim \text{InvWishart}(\nu_0, \boldsymbol{\Lambda_0^{-1}}) \tag{16}$$

$$\boldsymbol{\mu}|\boldsymbol{\Sigma} \sim \text{MVNormal}(\boldsymbol{\mu_0}, \boldsymbol{\Sigma}/\kappa_0) \tag{17}$$

This would add two additional sampling stages to the algorithm (see Gelman et al. (2014)):

1) Sample $\boldsymbol{\Sigma}$. Define the sufficient statistics, $\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x_i}$ and $\boldsymbol{S} = \sum_{i=1}^{N}(\boldsymbol{x_i} - \bar{\boldsymbol{x}})(\boldsymbol{x_i} - \bar{\boldsymbol{x}})^T$. Then,

$$\boldsymbol{\Sigma} \sim \text{InvWishart}(\nu_N, \boldsymbol{\Lambda_N^{-1}}), \tag{18}$$

where $\nu_N = \nu_0 + N$ and $\boldsymbol{\Lambda_N} = \boldsymbol{\Lambda_0} + \boldsymbol{S} + \frac{\kappa_0 N}{\kappa_0 + N}(\bar{\boldsymbol{x}} - \boldsymbol{\mu_0})(\bar{\boldsymbol{x}} - \boldsymbol{\mu_0})^T$.

2) Sample $\boldsymbol{\mu_k}|\boldsymbol{\Sigma_k}$ for $k = 1, ..., K^*$. Using the same defined terms in the previous step,

$$\boldsymbol{\mu}|\boldsymbol{\Sigma} \sim \text{MVNormal}(\boldsymbol{\mu_N}, \boldsymbol{\Sigma}/\kappa_N), \tag{19}$$

where $\boldsymbol{\mu_N} = \frac{\kappa_0}{\kappa_0+N}\boldsymbol{\mu_0} + \frac{N}{\kappa_0+N}\bar{\boldsymbol{x}}$ and $\kappa_N = \kappa_0 + N$.

Different specified distributions would require different sampling procedures. In a supplementary document, Royle (2009) considers other types of covariate distributions and found some variations in the inference. Unfortunately, knowing what distribution to specify can be difficult as rarely would we know this distribution. Deciding on a proposal distribution based upon the observed distribution can be also dangerous or misleading as the observed distribution is a truncation of the true distribution. Not only are the missing covariates not missing at random but we do not even know how many are missing. As we will shown in section **??**, mispecifying the distribution can lead to inaccurate estimations.

3.3.2 Conditional Likelihood

One approach to this problem is to estimate the coefficients using conditional maximum likelihood, where the likelihood function to be maximized is conditioned on the probability an individual is observed at least once (Alho, 1990; Huggins, 1989). This avoids the necessity of specifying a distribution for the covariates but can lead to unstable results when the observed distribution of the covariates is dissimilar to the population distribution of the covariates (Tilling and Sterne, 1999). An article by Yee et al. (2015) shows an easy way to implement the technique using the `VGAM` package in `R`. The technique works in two stages. First, use generalized linear models with a positive Bernoulli family to estimate the the coefficients. This allows us to obtain fitted values for the probability that each individual in the dataset is missing. Second, use those fitted values in the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) to estimate the population size, N. One could take a Bayesian approach to the logistic regression and assign prior distributions to the parameters and estimate the population in a similar way. We derive

equations for finding the maximum a posteriori (MAP) estimate in Appendix 8 using gradient ascent.

3.3.3 Proposed Extension 1: Nonparametric Distributions

We propose an alternative approach which is to specify a flexible distribution to the covariates such as a Dirichlet process mixture of normal distributions. Using stick-breaking priors (Ishwaran and James, 2001), this nonparametric approach fits a potentially infinite number of multivariate normal distributions to the data. The implementation is adapted from Gelman et al. (2014). This model introduces a latent variable, z, which is a latent parameter determining the mean and covariance matrix from which the observed covariate, $\boldsymbol{x_i}$, is drawn. The generative process for each observation's covariate can be summarized,

$$\boldsymbol{x_i}|z_i \overset{ind}{\sim} \text{MVNormal}(\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \text{ for } i = 1, ..., N \tag{20}$$

$$z_i \overset{iid}{\sim} \text{Discrete}(\{1, 2, ...\}, (\pi_1, \pi_2, ...)) \text{ for } i = 1, ..., N. \tag{21}$$

Using this generative scheme adds five additional sampling stages to the algorithm presented in subsection 3.2. The full details including the prior specifications are provided in Appendix 9. While the number of latent classes is infinite, we approximate this problem by truncating the number of latent classes, K^*, and solving this finite-dimensional problem. This upper bound, K^*, should not be thought of as a parameter as it should have no impact on the estimation process as long as the value is set sufficiently large enough. It should be noted that the normal distribution specified in subsection 3.3.1 is a special case of this specification where $K^* = 1$.

3.3.4 Future Work: Dirichlet Process

In the previous subsections, we considered a single multivariate normal distribution but also an infinite mixture of normal distributions with membership determined through a stick-breaking process. We will show in section **??** that the assumption of a single normal may lead to bias in the estimation process when the actual distribution is not normal. The infinite mixture of normal distributions tends to perform better, but struggles when the distribution is discrete or is far from normally distributed. An alternative approach would be using the empirical distribution of the covariate. This has a similar intuition as the mechanic behind the Horvitz-Thompson estimator in the frequentist approach where each covariate essentially gets magnified based on its estimated probability of being missing. This approach could be made more general by using Dirichlet process mixtures. As shown in Gelman et al. (2014), a sufficiently large concentration parameter, $\alpha \to \infty$, converges to the empirical distribution.

3.4 Conditional Independence and Unobserved Heterogeneity

The model construction of section 3 assumes conditional independence based on the covariates. In other words, given the information provided by the covariates, the probability of capture on one list is unaffected by another list. If the assumption does not hold, it may lead to biased parameter estimates. One reason the conditional independence may be violated is because the capture probabilities upon a list are directly related to the probability of being on another list, i.e., the assumption of list dependency. This may occur if, for example, one list uses another list as a reference or data is shared between various documentation projects (Manrique-Vallier et al., 2020). This would lead to heavy positive dependence between these two lists. While this is a serious issue, we assume the lists used in the analysis are collected independently.

A second reason conditional independence may be violated is that underlying heterogeneity

exists within the population that is not fully accounted for by covariates. Recall, the example cited in subsection 2.1 of individual heterogeneity masquerading as list dependency in the study analyzing extrajudicial killings during the Guatemalan Civil War (Ball et al., 2000). Researchers found that people who were part of Catholic religious communities were more likely to trust Catholic researchers with their stories than with NGO researchers associated with the political left. Similarly, people located in areas associated with the rebel groups were more likely to do the opposite. If this trait is unobserved and not taken into account, it will result in biased $\boldsymbol{\beta_j}$ coefficients. Further, because of the biased coefficients, the probabilities of capture and the estimate for the population size will be biased as well (see section **??**).

3.4.1 Proposed Extension 2: Modelling the Unobserved Heterogeneity with Latent Classes

We view the problem of unobserved heterogeneity through the lens of a missing covariates problem. While there are multiple ways one could implement additional heterogeneity, we choose to add an indicator vector, $\boldsymbol{\omega_i}$, that indicates membership to one of H_ω latent groups with probability, $\boldsymbol{\phi_\omega} \sim \text{Dirichlet}(\boldsymbol{\alpha_\omega})$. For simplicity, assume a hyperparameter specification with each value assigned the same value, α_ω. Larger values of this hyperparameter put more weight on the prior and less on the data.

This leads to three new types of parameters to sample: $\boldsymbol{\omega_i}$, $\boldsymbol{\phi_\omega}$, and $\boldsymbol{\beta_\omega}$. The model is constructed such that each latent group has an additional intercept affecting the probability of capture on each list. Equation 8 then becomes

$$\lambda_{ij} = \sigma(\boldsymbol{x_i}^T\boldsymbol{\beta_j} + \boldsymbol{\omega_i}^T\boldsymbol{\beta_\omega}) = \frac{1}{1 + e^{-(\boldsymbol{x_i}^T\boldsymbol{\beta_j} + \boldsymbol{\omega_i}^T\boldsymbol{\beta_\omega}))}}, \tag{22}$$

3.4.2 Proposed Extension 2: Updating the Estimation

Since we have three new parameters: $\boldsymbol{\omega_i}$, $\boldsymbol{\phi_\omega}$, and $\boldsymbol{\beta_\omega}$; we might expect to need just three new sampling methods. Unfortunately, complications once again arise from the missing covariates. If the individual is observed, the latent group membership, $\boldsymbol{\omega_i}$, can be sampled with corresponding discrete probability,

$$p(\boldsymbol{\omega_i}|\boldsymbol{\phi_\omega}, N, \mathcal{Y}, \boldsymbol{\beta}, \boldsymbol{X}) \propto \phi_\omega \prod_{j=1}^{J} \lambda_{ij}^{y_{ij}} (1-\lambda_{ij})^{1-y_{ij}}. \tag{23}$$

For the individuals that are not observed, we draw the latent class membership during the missing covariate imputation stage. Recall, because of the complications mentioned in subsection 3.2.2, the Gibbs sampler requires N and $\mathcal{X}_{mis}$ to be sampled simultaneously. In order to implement the sampling procedure of the missing covariates, $\mathcal{X}_{mis}$, the latent class membership, $\boldsymbol{\omega_i}$ of each unobserved individual must be drawn initially as well from its subpopulation with probability $\boldsymbol{\phi_\omega}$.

The $\boldsymbol{\phi_\omega}$ are sampled according to,

$$\boldsymbol{\phi_\omega} \sim \text{Dirichlet}(\alpha_\omega + n_{\omega=1}, \alpha_\omega + n_{\omega=2}, ..., \alpha_\omega + n_{\omega=H_\omega}), \tag{24}$$

where n_ω is the number of individuals belonging to each latent class membership and α_ω is a hyperparameter which can be thought of as a prior sample size. We set $\alpha_\omega = 1$. Lastly, the additional coefficient vector, $\boldsymbol{\beta_\omega}$, can be sampled in the same manner and simultaneously with the other coefficients, $\boldsymbol{\beta_j}$, when conditioned on $\boldsymbol{\phi_\omega}$.

3.4.3 Future Work: Implementing Stick-Breaking Priors for Latent Classes

The current setup uses a finite number of latent classes with a prior specification of a Dirichlet($1/H_\omega$,...,$1/H_\omega$).While it may be reasonable in some cases to know the number of

latent classes, it may be advantageous to utilize the stick-breaking prior. Of course, with a sufficiently large H_ω the current construction will approximate the solution under the stick-breaking prior. Nevertheless, the current construction uses the concentration parameter, α_ω, as a hyperparameter. Adding some flexibility by placing a prior on α_ω could prove beneficial.

3.5 Proposed Extension 3: Covariate Selection

The covariates play an important role in how they affect the estimated capture probabilities that ultimately influence the estimate of the population size N. If many covariates are present, we need a method for determining which covariates should be used along with possible interactions, which would relax the independence assumption between covariates. Even in the presence of a single covariate, how that covariate is used can have a major effect on the inference process. Perhaps some sort of non-linear transformation such as the log would give a better linear fit. A couple of strategies come to mind. First, we could consider some form of model averaging where multiple models are considered with various combinations of covariates. Alternatively, we could implement a prior that induces variables selection such as the horseshoe prior or spike and slab prior. A third possible solution is to perform some sort of post-hoc analysis between various fits and compute the Bayes factor.

4 Simulation Analysis

In this section, we run numerous simulations covering a number of different types of situations including varying levels of list dependency, sizes of population, non-normal covariate distributions, and unobservable heterogeneity. The primary objective is to examine the results of the Bayesian Logistic Regression Capture-Recapture (BLRCR) model. Along with the BLRCR algorithm, we compare our results using four other capture-recapture algorithms. The first algorithm we use is conditional maximum likelihood logistic regression (cMLCR) which is implemented using the `VGAM` package in `R` (see subsection 3.3). Instead of using the asymptotic estimates for the standard error and assuming normality, we use a semiparametric bootstrap for the confidence intervals (Zwane and van der Heijden, 2003). Second, we implement the ubiquitous hierarchical log-linear (Log Linear) modelling technique (Fienberg, 1972) using the the `Rcapture` package in `R`. Keep in mind that this approach does not use covariates, but attempts to model list dependency directly through list interactions. Since the approach is hierarchical with 2^J different model constructions, all are calculated and the one with the lowest BIC is selected. Third, we use the Bayesian Non-Parametric Latent-Class Capture-Recapture (LCMCR) algorithm in Manrique-Vallier (2016) which is another technique that does not allow covariates but uses a Bayesian nonparametric approach to account for unobserved heterogeneity (see Section 10: Appendix C for a summary of this approach). The fourth technique is a simple independence model (Independent) where it is assumed there is no list or individual heterogeneity. Conveniently, when the number of latent classes is set equal to one, the LCMCR model collapses into an independence model, effectively giving a Bayesian independence sampler (Independent).

We estimate N using the BLRCR model under three different specifications. First, we consider a single multivariate normal to describe the covariate distribution ($K = 1$) and no hidden heterogeneity (H_ω=1). Second, we allow for a mixture of multivariate normal

distributions under the stick-breaking prior with a sufficiently large number of classes, $K = 20$, but still do not allow for additional unobserved heterogeneity. The third specification is similar to the second but now allows hidden heterogeneity with up to $H_\omega = 20$ different latent intercepts. Also, the BLRCR model requires a prior mean and covariance for the coefficients, which are set to

$$\boldsymbol{b} = \mathbf{0} \text{ and } \boldsymbol{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

In addition to priors on the coefficients, BLRCR and LCMCR require additional hyperparameter specifications for the mixture distributions which can be found in Table 1. The number of samples in each simulation for each method is set to 10,000 unless otherwise stated.

	Method	
Hyperparameter	SP-BLRCR	LCMCR
a	0.25	0.25
b	0.25	0.25
ν_0	3	
κ_0	1	
$\boldsymbol{\mu_0}$	(0,0)	
$\boldsymbol{\Lambda_0}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	

Table 1: Hyperparameter specifications for the SP-BLRCR and LCMCR methodologies.

The following subsections run simulations on data generated from different characteristics including levels of list dependency (subsection 4.2), size of population (subsection 4.1), various covariate distributions (subsection 4.3), and unobservable heterogeneity (subsection 4.4). The objective is to compare and contrast the BLRCR model with the other approaches described above. For each algorithm a point estimate of the population size, N, is computed. We have some options for what to use as the point estimate, $\hat{N}$, for the

MCMC algorithms, but we elect to use median of the sampled posterior. In order to get a measure on the precision and accuracy model, we also compute the 95% confidence/credible interval.

In each subsection a table with results can be found. The column $N\%$ computes the average $\hat{N}$ across the simulations and takes it as a percentage of N. Hence, a score of 1.000 would indicate an unbiased estimate. In addition, we consider the accuracy of the point estimate by computing the mean squared error (MSE) of the simulated $\hat{N}$. We also look to the accuracy of the 95% interval estimate by checking whether N fell inside that interval (CI%). Of course, the precision of the interval must also be considered so we computed the average of the simulated 95% confidence/credible interval widths as a percentage of N (CI Width). Ideally, we would want a model with a small interval width (high precision) but maintains the ability to find the true population size often (high accuracy).

4.1 Simulations with Different Sized Populations

In this subsection, we examine the effects of differing population sizes on the estimations with two objectives to evaluate in mind: performance and consistency. For the most part, we will see that all models will become less biased as the population size increases (except assuming independence). We will also see that using the BLRCR will tend to have a lower MSE than the cMLCR approach, especially with smaller population sizes.

Data was simulated with varying population sizes, N =[200,500,1000,2000,5000,10000], with the results summarized in Table 2. For consistency, all datasets were contructed using the coefficients yielding "moderate" dependency and with two standard normal covariates. Before examining the results, it should be pointed out that there were considerable issues with using the log linear approach with the smaller sample sizes. While a point estimate was always able to be obtained, often times, the `Rcapture` package would simply report a lower bound on the upper limit of the confidence interval. Instead of making a decision on

whether to use that bound as the upper limit, we simply report the interval width as not available (NA).

With this consideration in mind, we highlight a few interesting observations. Most importantly, the BLRCR and cMLCR methods tended to outperform the other methods in terms of mean squared error (MSE), 95 % credible interval width (CI width), and the 95% credible interval capture percentage (CI %). There were a handful of occasions where that was not the case but appears to be attributed to the choice of simulation parameters. Noteably, the independence model outperformed all other models when $N = 1000$, but as the population grew, it became substantially worse and is actually the worst option when $N = 10000$. Additional simulations (not shown) performed with different sets of coefficients revealed the independence model to be a substantially worse choice than BLRCR regardless of population size.

A comparison of the three BLRCR methods and cMLCR reveal relatively similar performances across population sizes. Keep in mind, the simulation used in Table 4 simply used a single multivariate normal with no heterogeneity. The additional noise of assuming an infinite mixture of normal distributions and/or heterogeneity in capture probabilities seems to have had little to no impact on the estimation.Figure 1 shows a correlation plot between the 100 estimates for N using the three BLRCR methods and cMLCR with the true population size set to $N = 2000$. Notice, the estimates are highly correlated, and BLRCR tends to return smaller estimated values than cMLCR. An important result of Alho (1990), is that the cMLCR estimator is consistent but may be biased with small sample sizes. From the simulations, it certainly appears this is true for N=200, but the level of bias quickly disappears with N=500 and above. Similarly, the BLRCR seems to share this quality of consistency but the bias is greater albeit the opposite direction. Nevertheless, as we would expect with Bayesian methods, we are trading some bias for a reduction in variance. The mean squared error (MSE) tends to be smaller than the other methods.

N	Method	N%	MSE	CI Width	CI %
200	BLRCR($K = 1$, $H_\omega = 1$)	0.887	28.8	0.374	78.0
	BLRCR($K = 20$, $H_\omega = 1$)	0.883	29.4	0.375	78.0
	BLRCR($K = 20$, $H_\omega = 20$)	0.914	26.4	0.459	92.0
	cMLCR	1.057	36.7	0.635	96.0
	Log Linear (BIC)	2.316	472.6	NA	23.0
	LCMCR	0.854	35.7	0.434	74.0
	Independent	1.004	22.0	0.481	97.0
500	BLRCR($K = 1$, $H_\omega = 1$)	0.940	44.0	0.271	91.0
	BLRCR($K = 20$, $H_\omega = 20$)	0.943	42.7	0.280	90.0
	BLRCR($K = 20$, $H_\omega = 20$)	0.966	37.6	0.342	96.0
	cMLCR	1.010	41.0	0.340	95.0
	Log Linear (BIC)	2.084	690.8	1.826	1.0
	LCMCR	0.840	84.1	0.253	43.0
	Independent	1.017	35.4	0.306	96.0
1000	BLRCR($K = 1$, $H_\omega = 1$)	0.972	58.3	0.210	93.0
	BLRCR($K =$, $H_\omega = 20$)	0.976	56.9	0.217	94.0
	BLRCR($K = 20$, $H_\omega = 20$)	0.991	57.9	0.267	97.0
	cMLCR	1.008	60.1	0.235	95.0
	Log Linear (BIC)	1.808	1106.6	0.957	0.0
	LCMCR	0.852	154.5	0.190	20.0
	Independent	1.033	62.8	0.222	92.0
2000	BLRCR($K = 1$, $H_\omega = 1$)	0.983	82.5	0.156	94.0
	BLRCR($K = 20$, $H_\omega = 1$)	0.988	82.6	0.162	95.0
	BLRCR($K = 20$, $H_\omega = 20$)	0.997	84.9	0.204	99.0
	cMLCR	1.002	87.1	0.163	92.0
	Log Linear (BIC)	1.399	1669.3	0.458	0.0
	LCMCR	0.851	303.4	0.148	11.0
	Independent	1.035	103.1	0.157	90.0
5000	BLRCR($K = 1$, $H_\omega = 1$)	0.993	123.7	0.107	96.0
	BLRCR($K = 20$, $H_\omega = 1$)	0.997	120.2	0.112	98.0
	BLRCR($K = 20$, $H_\omega = 20$)	1.002	127.2	0.145	100.0
	cMLCR	1.001	122.1	0.102	96.0
	Log Linear (BIC)	1.024	1651.3	0.326	62.0
	LCMCR	0.898	763.4	0.277	37.0
	Independent	1.039	224.4	0.100	65.0
10000	BLRCR($K = 1$, $H_\omega = 1$)	0.996	174.0	0.084	97.0
	BLRCR($K = 20$, $H_\omega = 1$)	0.998	177.0	0.087	98.0
	BLRCR($K = 20$, $H_\omega = 20$)	0.997	197.0	0.109	99.0
	cMLCR	0.999	178.9	0.072	94.0
	Log Linear (BIC)	1.000	801.5	0.278	95.0
	LCMCR	0.957	1491.2	0.385	64.0
	Independent	1.039	420.8	0.071	35.0

Table 2: Results of 100 capture-recapture simulations per varying population sizes using "Moderate" list dependency and two standard normal covariates.

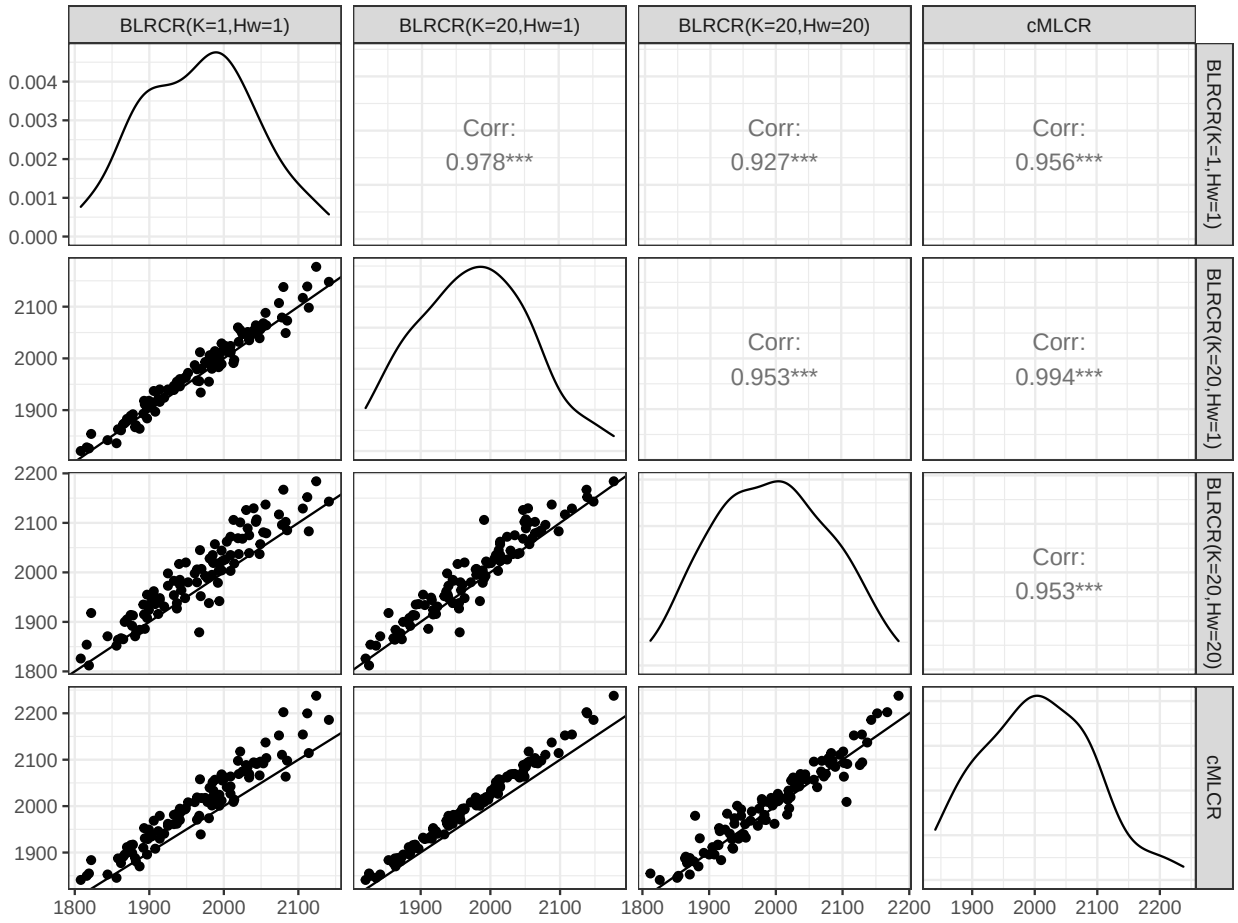

Figure 1: Correlation Plot of the 100 estimates for N using BLRCR and cMLCR when simulating from two independent standard normal distributions with moderate dependency between lists and a true population size of N=2000.

Overall, the BLRCR model is shown to be consistent. It also tends to outperform or equally perform the other methods in terms of MSE, CI width, and CI%.

4.2 Simulations with Varying Levels of List Dependency

The objective of this subsection is to evaluate the BLRCR and other models under varying coefficients that induce different levels of list dependency. All of this list dependency is really just heterogeneity in the capture probabilities that can be fully explained by the covariates. The models that do not use covariates (Log Linear, LCMCR, and Independent) will not or will struggle to detect the heterogeneity leading to biased results. On the other

hand, the BLRCR and cMLCR will perform well as they incorporate the covariates into their modelling.

We simulate data in accordance with Equation 6 using two covariates (H=2) and four lists ($J = 4$). The covariates are drawn from two independent standard normal distributions. Three sets of β coefficients are chosen that create "negative", "moderate", and "positive" dependency between the lists and are described in Table 3. To clarify, all of the lists regardless of the coefficients are conditionally independent given the covariates. We simulate an additional dataset with all of the slope coefficients set to 0, thereby creating lists with independent capture probabilities.

"Negative"				"Moderate"				"Positive"			
List (j)	β_{0j}	β_{1j}	β_{2j}	List (j)	β_{0j}	β_{1j}	β_{2j}	List (j)	β_{0j}	β_{1j}	β_{2j}
1	-2	-1	1	1	-2	-1	1	1	-2	-1	1
2	-2	1	-1	2	-2	1	-1	2	-2	-1	1
3	-2	1	1	3	-2	-1	1	3	-2	-1	1
4	-2	-1	-1	4	-2	1	-1	4	-2	-1	1

Table 3: Coefficients for Simulated Data

$\boldsymbol{\beta}$	Method	N%	MSE	CI Width	CI %
"Moderate"	BLRCR($K = 1$, $H_\omega = 1$)	0.983	82.5	0.156	94.0
	BLRCR($K = 20$, $H_\omega = 1$)	0.988	82.6	0.162	95.0
	BLRCR($K = 20$, $H_\omega = 20$)	0.997	84.9	0.204	99.0
	cMLCR	1.002	87.1	0.163	92.0
	Log Linear (BIC)	1.399	1669.3	0.458	0.0
	LCMCR	0.851	303.4	0.148	11.0
	Independent	1.035	103.1	0.157	90.0
"Negative"	BLRCR($K = 1$, $H_\omega = 1$)	0.987	79.0	0.149	97.0
	BLRCR($K = 20$, $H_\omega = 1$)	0.990	76.3	0.151	100.0
	BLRCR($K = 20$, $H_\omega = 20$)	1.006	80.5	0.189	99.0
	cMLCR	1.000	77.1	0.150	99.0
	Log Linear (BIC)	0.999	81.0	0.159	98.0
	LCMCR	1.008	99.3	0.284	98.0
	Independent	1.208	426.7	0.206	0.0
"Positive"	BLRCR($K = 1$, $H_\omega = 1$)	0.942	188.0	0.350	91.0
	BLRCR($K = 20$, $H_\omega = 1$)	0.952	168.1	0.354	93.0
	BLRCR($K = 20$, $H_\omega = 20$)	0.978	158.0	0.385	97.0
	cMLCR	1.025	207.3	0.389	94.0
	Log Linear (BIC)	0.899	301.4	0.322	77.0
	LCMCR	0.699	608.1	0.236	4.0
	Independent	0.566	868.4	0.048	0.0
"Independent"	BLRCR($K = 1$, $H_\omega = 1$)	0.981	120.2	0.247	96.0
	BLRCR($K = 20$, $H_\omega = 1$)	0.980	119.2	0.247	96.0
	BLRCR($K = 20$, $H_\omega = 20$)	1.009	128.6	0.370	100.0
	cMLCR	1.011	126.6	0.261	97.0
	Log Linear (BIC)	1.002	121.5	0.256	98.0
	LCMCR	0.993	115.7	0.260	99.0
	Independent	0.989	115.9	0.247	97.0

Table 4: Results of 100 capture-recapture simulations per coefficient set using $N = 2000$ with "Moderate", "Negative", "Positive", and "Independent" list dependency with two standard normal covariates.

Table 4 shows the results from 100 simulated datasets for each set of coefficients, β. While all of the models performed well when the slope coefficients were set to 0, i.e. “Independent”, the methods using covariates (BLRCR and cMLRCR) showed substantially less bias than the other methods when coefficients impacted the capture probability, i.e. “Moderate”, “Negative”, and “Positive”. This illustrates the importance of including covariates in the estimation process. Nevertheless, the two methods that account

for unknown heterogeneity, log linear and LCMCR, still performed remarkably well when the list dependency was set to be "negative." Unfortunately, "positively" induced list dependency led to a substantial decline in performance.

4.3 Simulations using Different Covariate Distributions

In this subsection, we examine the impact of different covariate distributions and how they impact the BLRCR model's estimation. Recall, the BLRCR requires the specification of a covariate distribution, whereas the cMLCR does not. Hence, as we will see below, a misspecification of the covariate distribution can lead to bias in the inference. This bias will be reduced when using a more flexible covariate distribution like the infinite mixture of normal distributions (K=20).

Table 5 presents the results of 100 simulations for three different sets of covariate distributions. The first set of simulations uses the two standard normal distributions seen in the previous subsections. While the cMLCR algorithm is unbiased, all three specifications of the BLRCR algorithm yield a lower MSE. Since the distribution of the covariates is actually normal, it is not surprising that the BLRCR performs well with these simulations. On the other hand, we would expect the second set of covariates, two independent chi-square(1) distributions, to be particularly challenging. The chi-square distribution only has probability mass for nonnegative values, which creates an abrupt cutoff at 0. The third set of covariates includes two different Gamma distributions, Gamma(1,1) and Gamma(3,1). This set of covariates cuts off abruptly on one axis at 0, but not the other. We can therefore think of the three sets of covariates as "normal", "not normal", and "near/approximately normal", respectively.

Distribution	Method	N%	MSE	CI Width	CI %
Normal(0,1)	BLRCR($K = 1$, $H_\omega = 1$)	0.983	82.5	0.156	94.0
Normal(0,1)	BLRCR($K = 20$, $H_\omega = 1$)	0.988	82.6	0.162	95.0
	BLRCR($K = 20$, $H_\omega = 20$)	0.997	84.9	0.204	99.0
	cMLCR	1.002	87.1	0.163	92.0
	Log Linear (BIC)	1.399	1669.3	0.458	0.0
	LCMCR	0.851	303.4	0.148	11.0
	Independent	1.035	103.1	0.157	90.0
Chi-Square(1)	BLRCR($K = 1$, $H_\omega = 1$)	1.656	1383.4	0.950	0.0
Chi-Square(1)	BLRCR($K = 20$, $H_\omega = 1$)	0.967	117.6	0.178	89.0
	BLRCR($K = 20$, $H_\omega = 20$)	0.986	148.5	0.241	90.0
	cMLCR	1.008	104.8	0.183	92.0
	Log Linear (BIC)	0.892	973.3	0.217	6.0
	LCMCR	0.774	467.0	0.229	31.0
	Independent	0.924	162.8	0.118	32.0
Gamma(1,1)	BLRCR($K = 1$,$H_\omega = 1$)	1.020	90.7	0.156	94.0
Gamma(3,1)	BLRCR($K = 20$, $H_\omega = 1$)	0.986	63.4	0.132	99.0
	BLRCR($K = 20$,$H_\omega = 20$)	0.996	69.6	0.170	100.0
	cMLCR	0.999	55.3	0.123	96.0
	Log Linear (BIC)	1.247	1443.3	0.409	0.0
	LCMCR	0.852	299.3	0.106	7.0
	Independent	0.863	276.6	0.054	0.0

Table 5: Results of capture-recapture algorithms with simulations using different covariate distributions with "moderate" dependency between lists and N=2000.

As we saw in the previous section, the BLRCR algorithm handles the normally distributed covariates well. For the chi-square distributed covariates, using a single multivariate normal for the covariates results in poor performance with with an approximate bias of 1.656% of N when $N = 2000$. Even though the covariates are not normally distributed, modeling them as a mixture of normal distributions results in a substantial reduction in the bias for N. Figure 2 illustrates why this may be the case through a partial plotting of the simulated covariates. In all plots, the black dots represent the individuals that are captured at least once in a list. The blue dots represent the individuals that are missing. In the first and second plot in each row, the individuals that are missing are simulated using the BLRCR($K = 1$, $H_\omega = 1$) and BLRCR($K = 20$, $H_\omega = 1$) algorithms, respectively. The third panel shows the true missing individuals that are unknown to the algorithm.

With the chi-square distribution, there is no good way to fit a single normal variable that well represents the space. As a result, many missing covariates are populated into low probability density areas resulting in an overestimate of the missing covariates. On the other hand, the BLRCR($K = 20$, $H_\omega = 1$) with it's less rigid covariate assumption, populates the space much better.

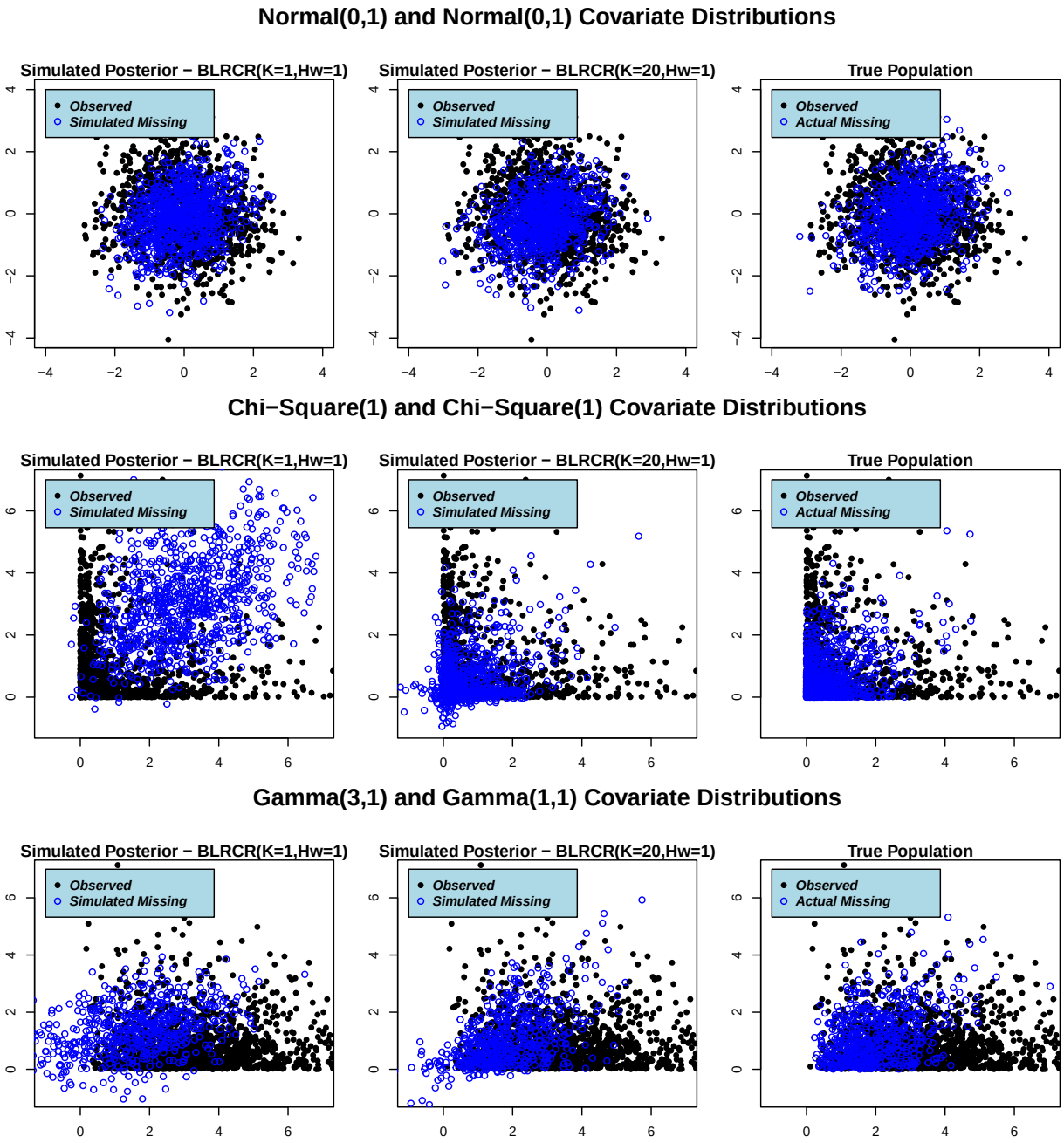

Figure 2: Posterior Distribution of X when simulating various covariate distributions with moderate dependency between lists and N=2000.

The parameters of the covariate distribution are sampled based on augmented covariates, not just the observed covariates. This results in uncertainty regarding the ability for the algorithm to correctly identify the covariate distribution. To put the algorithm to the test, we simulate an example with a mixture of three multivariate normal distributions. We use

a population of size 2000 and the "moderate" coefficients. The results of the 100 simulations can be found in Table 6. Despite on average nearly 50% of observations being missing, the algorithm when allowing for up to $K = 20$ mixture normal distributions performs well. Of course, the cMLCR outperforms the algorithm in terms of MSE, but the BLRCR (without also trying to account for heterogeneity) is more precise with a smaller average credible interal width.

Distribution	Method	N%	MSE	CI Width	CI %
Mixture Normal	BLRCR($K = 1$,$H_\omega = 1$)	0.931	160.1	0.163	67.0
	BLRCR($K = 20$,$H_\omega = 1$)	0.977	97.9	0.183	96.0
	BLRCR($K = 20$,$H_\omega = 20$)	0.989	96.3	0.228	99.0
	cMLCR	1.003	90.3	0.191	96.0
	Log Linear (BIC)	1.331	1331.4	0.427	0.0
	LCMCR	0.851	305.5	0.218	25.0
	Independent	1.007	80.0	0.173	96.0

Table 6: Results of capture-recapture algorithms with simulations using different covariate distributions with "moderate" dependency between lists.

A look at the simulated posterior of the covariates in Figure 3 shows once again the benefits of using the mixture of normal distributions. Using a single multivariate normal results in an imputation of covariates in low probability mass spaces, especially between the lower two mixtures. When a mixture of normal distributions is used, the imputation of the covariates is fairly consistent with the true distribution.

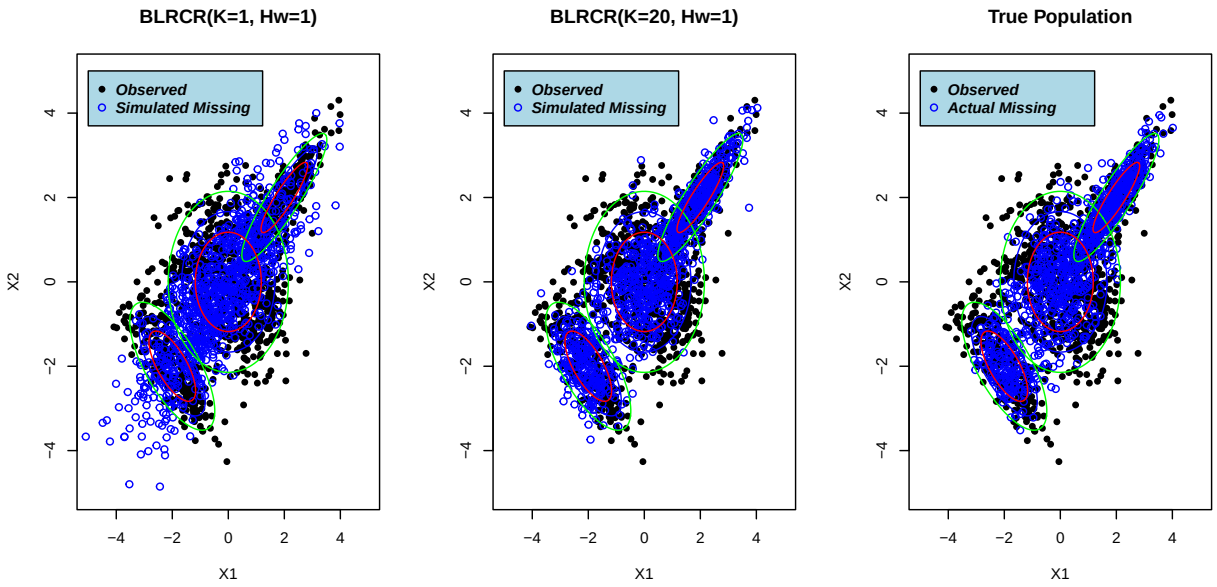

Figure 3: Posterior Distribution of the missing X values using algorithm SP-BLRCR.

From this subsection, we saw that depending on the level of the covariate distribution misspecification, the BLRCR model may perform poorly. However, using a non-parametric distribution like the mixture of normal distributions considerably reduced the bias.

4.4 Simulations with Unobserved Heterogeneity

The final set of simulations is to demonstrate the importance of accounting for both the observed and unobserved heterogeneity. While the cMLCR model has performed well in the prior sections, it has no method of detecting unobservable heterogeneity and hence produces biased estimation. Interestingly, a model like the LCMCR, which is designed to account for unobservable heterogeneity, performs much better, but still struggles as considerable information can be gained by including the covariates. The BLRCR model when extended to include latent intercepts ($H_\omega > 1$) can utilize the covariates to detect the observable heterogeneity, but also accounts for the additional unobservable heterogeneity. As a result, it is the only model that performs well in this subsection.

Data is simulated for three lists (J=3) with one standard normally distributed observed

covariate (H=1) and one unobserved covariate indicating membership to a latent group with probability 0.35. The coefficients used to simulate the data can be found in Table 7. Using these coefficients induces positive dependency between lists 1 and 2, but negative dependency between list 3 and the other two lists. If class membership were known, the list probabilities would still be conditionally independent; however, since these covariates are unobserved, the list probabilities are no longer conditionally independent given the observed data.

List (j)	β_{0j}	β_{1j}	$\beta_{\omega j}$
1	-2.5	-1.5	3.0
2	-2.5	-1.5	3.0
3	0.5	-1.5	-3.0

Table 7: Coefficients for Heterogeneity Simulated Data

Table 8 shows the results of 100 simulations on four different population sizes. It becomes obvious that not accounting for unobserved heterogeneity results in biased estimates for both the BLRCR and cMLCR models. Recall, the truth with these simulations is there exists two latent classes. Notice the BLRCR model with the number of hidden classes set at $H_\omega = 2$ performs the best in terms of both accuracy and precision. In a real setting, the number of hidden classes would almost certainly be unknown so we set a sufficiently large value $H_\omega = 20$. While the model doesn't perform quite as well as the aforementioned setting, it dramatically outperforms the methods that do not take unobserved heterogeneity into account.

For situations where observed and unobserved heterogeneity exist, we need a method that accounts for both. It should be noted that the LCMCR model, a methodology that doesn't use covariates, is outperforming the methods that do but assume conditional independence. The LCMCR model is designed to account for unobserved heterogeneity, and since the unobserved heterogeneity plays a substantial role in the capture probability for these simulations, it performs reasonably well. Of course, as we saw in the previous subsections'

simulations, the LCMCR does not perform as well as the other methods when most of the heterogeneity can be explained by the covariates.

To further explore the effects of hidden heterogeneity and its detectability, we ran 100 simulations with different coefficients and group percentages. Using the same coefficients in Table 7, but adjusting the absolute value of the coefficients, β_{ω_j}, we created scenarious that depict different strength levels of heterogeneity. Trivially, if the coefficient is 0, there is no unobserved heterogeneity in the capture probabilities. In this situation the algorithm is simply detecting noise that it is mistaking for heterogeneity. On the other hand, when the absolute value of the $\beta_{\omega j}$ coefficients are set to 5, there is very strong heterogeneity in the capture probabilities for the two groups. The top plot in Figure 4 shows the mean square error (MSE) of the 100 simulations' posterior median of N. A close examination of the plot reveals an overall decrease in the MSE as the heterogeneity strengthens. This is not surprising as the model is attempting to account for heterogeneity, but if it cannot detect the heterogeneity, it will induce bias. When the heterogeneity is stronger, the model is more likely to detect this heterogeneity and account for it properly.

N	Method	N%	MSE	CI Width	CI %
1000	BLRCR($K = 1$, $H_\omega = 1$)	1.192	209.0	0.386	25.0
	BLRCR($K = 20$, $H_\omega = 1$)	1.192	207.6	0.382	24.0
	BLRCR($K = 20$, $H_\omega = 2$)	0.997	57.5	0.334	100.0
	BLRCR($K = 20$, $H_\omega = 20$)	0.986	67.9	0.385	100.0
	cMLCR	1.231	250.8	0.390	25.0
	Log Linear (BIC)	1.048	94.7	0.199	64.0
	LCMCR	0.851	156.0	0.414	93.0
	Independent	0.875	128.2	0.095	1.0
2000	BLRCR($K = 1$, $H_\omega = 1$)	1.210	438.8	0.302	1.0
	BLRCR($K = 20$, $H_\omega = 1$)	1.205	428.4	0.295	1.0
	BLRCR($K = 20$, $H_\omega = 2$)	1.005	115.4	0.278	99.0
	BLRCR($K = 20$, $H_\omega = 20$)	0.994	130.0	0.316	98.0
	cMLCR	1.225	469.9	0.274	1.0
	Log Linear (BIC)	1.022	182.7	0.144	44.0
	LCMCR	0.858	301.0	0.388	93.0
	Independent	0.872	259.3	0.067	0.0
5000	BLRCR($K = 1$, $H_\omega = 1$)	1.229	1161.6	0.223	0.0
	BLRCR($K = 20$, $H_\omega = 1$)	1.222	1127.8	0.219	0.0
	BLRCR($K = 20$, $H_\omega = 2$)	1.010	217.2	0.208	100.0
	BLRCR($K = 20$, $H_\omega = 20$)	1.005	236.1	0.235	100.0
	cMLCR	1.230	1165.8	0.173	0.0
	Log Linear (BIC)	0.954	478.7	0.108	22.0
	LCMCR	0.869	700.0	0.340	85.0
	Independent	0.874	630.3	0.042	0.0
10000	BLRCR($K = 1$, $H_\omega = 1$)	1.224	2261.2	0.183	0.0
	BLRCR($K = 20$, $H_\omega = 1$)	1.219	2202.3	0.180	0.0
	BLRCR($K = 20$, $H_\omega = 2$)	1.005	359.1	0.164	99.0
	BLRCR($K = 20$, $H_\omega = 20$)	1.008	406.3	0.191	98.0
	cMLCR	1.221	2230.8	0.120	0.0
	Log Linear (BIC)	0.917	935.4	0.088	5.0
	LCMCR	0.869	1426.7	0.322	77.0
	Independent	0.872	1277.3	0.030	0.0

Table 8: Results of 100 capture-recapture simulations per algorithm using a standard normal distribution distribution for the known covariate and 0.35 probability of belonging to the latent class.

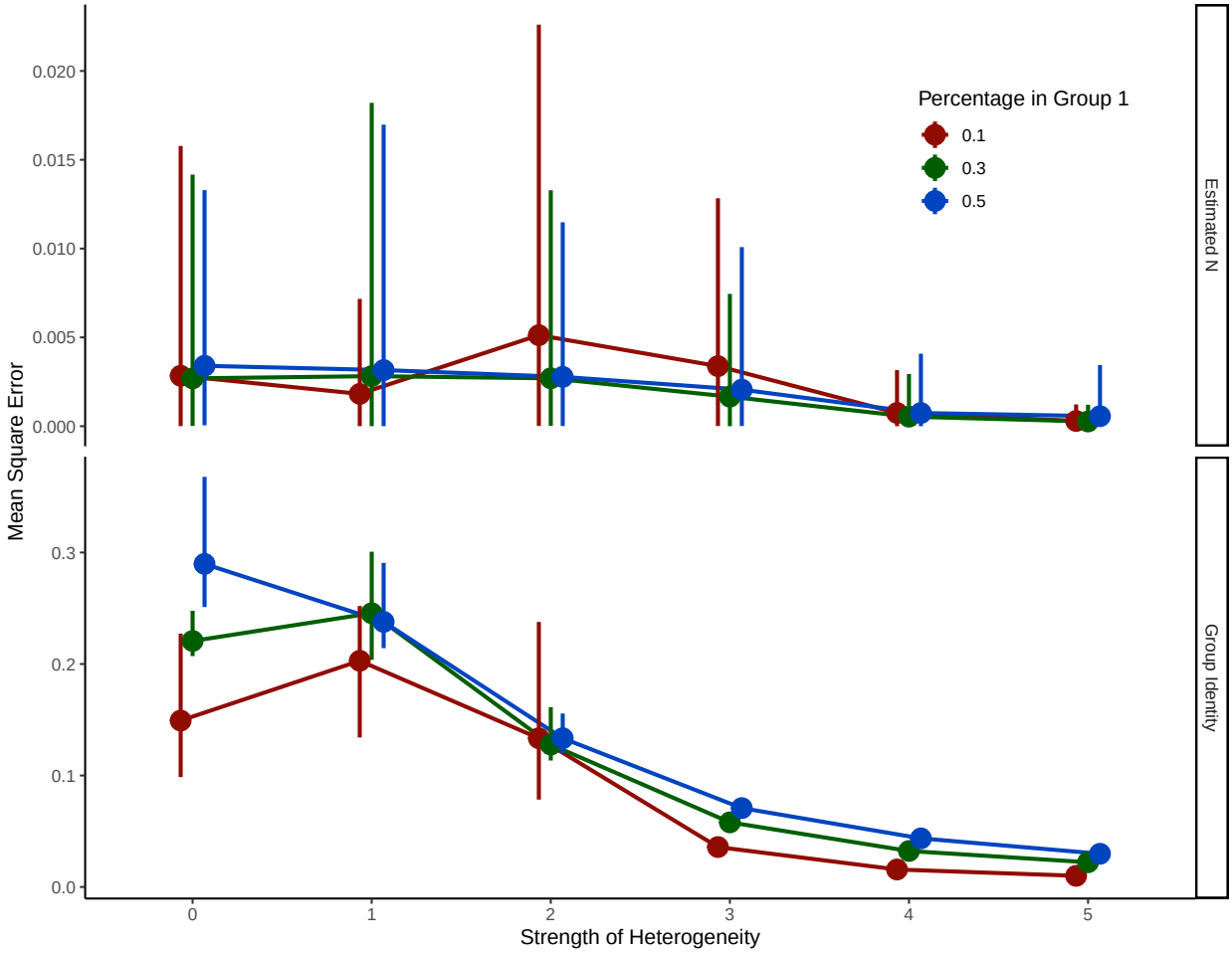

Figure 4: Effect of Strength of Heterogeneity on the Mean Square Error

The bottom plot in Figure 4 shows the MSE of the average posterior group identifier for the observed data. In other words, if the first observation actually belongs to the hidden group, ω_1=1, then we find the average number of times the algorithm placed the observation in the first group, $\bar{\omega}_1$. In the binary case, $H_\omega = 2$, the MSE can be computed as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\bar{\omega}_i - \omega_i)^2. \tag{25}$$

As the heterogeneity strengthens, the detectability of the unobserved groupings increases and the MSE drops considerably. In the case where the absolute value of $\beta_{\omega j}$ is set to 5, in many of the simulations between 90 and 95% of the observations are detected correctly in

over 90% of posterior samples.

Unobserved heterogeneity creates a trade off. On one hand, if the unobserved heterogeneity is weak, there will be relatively little bias in N; however, the ability to detect this heterogeneity decreases as well. On the other hand, When the unobserved heterogeneity is strong, the bias in N will be relatively larger, but the ability to detect and properly account for this heterogeneity increases.

5 Example 1 (probably superhero)

6 Example 2 (find something real even if we don't know the answer)

7 Conclusion

The objective of this dissertation project is to develop a Bayesian capture-recapture (CR) method that can utilize covariate information to understand the heterogeneity between individuals. Presented in this proposal is a framework for modelling capture-recapture with covariates. Specifically, we develop the Bayesian Logistic Regression Capture Recapture (BLRCR) that utilizes the covariate information directly in estimating individual capture rates. Further, we account for unobserved heterogeneity with the use of latent classes. This model is solved using an MCMC algorithm to approximate the posterior distribution of our population size, N. I propose three primary areas of interest that will be addressed as the focus of the dissertation: unobserved heterogeneity, modelling the covariate distribution, and covariate selection.

7.1 Unobserved Heterogeneity

This extension concerns the issue of unobserved heterogeneity described in subsection 3.4. To account for this heterogeneity, the algorithm detects latent groups and applies different intercepts in the linear term that defines the capture probability. Currently the number of latent classes is fixed and must be selected by the practitioner. An extension would be to include some method for determining the number of latent classes to include. While it may be reasonable in some cases to know the number of latent classes, an obvious approach would be to incorporate a stick-breaking prior on the number of latent classes.

7.2 Modelling the Covariate Distribution

A requirement of the algorithm is the necessity of specifying a distribution for the missing covariates, discussed in subsection 3.3. Typically this distribution is unknown so utilizing a distribution that is fairly flexible is ideal. In this proposal, we implemented the non-parametric approach of using an infinite mixture of normals with membership

determiend through a stick-breaking process (and additionally a single multivariate normal which is a special case). While the infinite mixture of normal distributions tends to perform well when the distribution is at least somewhat normal, it struggles when presented with certain non-normal distributions. Alternatively, I propose using Dirichlet process mixtures as described in Gelman et al. (2014). A sufficiently large concentration parameter, $\alpha \to \infty$, would be in essence the same thing as the empirical distribution.

7.3 Covariate Selection

If many covariates are present, we need a method for determining which covariates should be used while also considering interaction effects. Along these same lines, we need to consider how a variable affects the capture probability as some sort of transformation may be required. We suggest three possible routes including model averaging, variable selection through priors, or even post-hoc analysis like computing the Bayes factor.

7.4 Further Ideas and Futurework

The extensions mentioned above are of primary focus for the dissertation, but there are several other minor extensions that may be considered. A summary of all proposed extensions and future ideas can be found in Table 9. Extensions with priority equal to 1 will be addressed in the dissertation. Extensions with priority greater than 1 will be explored if time permits with lower values having higher priority. Extensions that do not get covered will be left for future work.

Extension	Description	Priority
Stick Breaking Priors	Instead of fixing the number of unobserved latent intercepts, we could use stick breaking priors.	1
Distribution for Covariates	Currently we implement a mixture of normals to represent the missing covariates. This is shown through simulation to work fairly well; however, often times we have non-continuous covariates or covariates that differ greatly from normality. We will explore the empirical distribution or a more generalized Dirichlet process for the covariates.	1
Covariate Selection	Currently, the model assumes the covariates to be used are known. Even with a single covariate, the relationship may be non-linear. One way to address this issue is through the use of polynomials or interactions of the covariates. Since this may cause the parameter space to get quite large, this could require a need for variable selection.	1
Create an R Package	We plan to develop an R package that allows practitioners the easy and fast ability to use the algorithm	1
Missing Covariates of Observed Data	Using the methodology described in this paper, the imputation of missing covariates should be naturally imputable.	2
The Probability of Missing	Currently we use a Monte Carlo within MCMC to find the probability of missing (see Bonner and Schofield (2014)). It may be advantageous to try different types of methods or even approximations like a Laplace Approximation (see Herliansyah et al. (2022))	2
Conditional Likelihood Newton's Method	Develop an algorithm to get the MAP estimate using Newton's Method. Could find confidence interval using bootstrap.	3

7.5 Current Status of Project and Timeline

From this proposal document, we can see that the basic framework and the Bayesian Logistic Regression (BLRCR) model have already been derived and studied. While considerable work has already been completed on modelling the unobserved heterogeneity (extension 1) and specifying the covariate distribution (extension 2), there is still some work left to be completed. As of now, little has been done in the area of covariate selection (extension 3), which will encompass most of the future effort. The current plan is to finish up with the first two extensions and then begin exploring methods for handling covariate selection.

The preliminary results of section **??** only utilize simulations. While simulations are an excellent tool to evaluate the success of an algorithm, the final project will use some real datasets.

The objective is to finish the dissertation project by May 2024.

8 Appendix A: Conditional Maximum Likelihood Estimation

Instead of using the full likelihood we replace it with the conditional likelihood, conditioned on each individual being observed once.

$$\begin{aligned} p(\mathcal{Y}|N, \boldsymbol{\beta}, \mathcal{X}_{obs}, \mathcal{X}_{mis}; \boldsymbol{y} \neq \mathbf{0}) =& \frac{p(\mathcal{Y}|N, \boldsymbol{\beta}, \mathcal{X}_{obs}, \mathcal{X}_{mis})}{P(\boldsymbol{y} \neq \mathbf{0})} \\ =& \prod_{i=1}^{N} \frac{\prod_{j=1}^{J} \lambda_{ij}^{y_{ij}} (1-\lambda_{ij})^{1-y_{ij}}}{1 - \prod_{j=1}^{J}(1-\lambda_{ij})} \end{aligned} \tag{26}$$

The posterior distribution can therefore be written as

$$\begin{aligned} p(N, \boldsymbol{\beta}, \mathcal{X}_{mis}|\mathcal{Y}, \mathcal{X}_{obs}) \propto & \left[\prod_{i=1}^{N} \frac{\prod_{j=1}^{J} \lambda_{ij}^{y_{ij}} (1-\lambda_{ij})^{1-y_{ij}}}{1 - \prod_{j=1}^{J}(1-\lambda_{ij})}\right] \times \left[\prod_{i=1}^{N} \boldsymbol{g}(\boldsymbol{\theta_h})\right] \times \left[\frac{1}{N}\right] \\ & \times \left[\prod_{j=1}^{J} \left(\frac{1}{2\pi}\right)^{H/2} |B|^{-1/2} e^{-\frac{1}{2}(\boldsymbol{b}-\boldsymbol{\beta_j})^T \boldsymbol{B}^{-1}(\boldsymbol{b}-\boldsymbol{\beta_j})}\right]. \end{aligned} \tag{27}$$

Instead of computing a MCMC sampler, we will instead compute the Maximum a posteriori (MAP) estimate for the coefficients, $\boldsymbol{\beta}_{MAP}$, using gradient ascent. Then, this estimate can be plugged into a Horvitz-Thompson estimator to find the N_{MAP}.

Taking into consideration only the parts of the posterior that depend on $\boldsymbol{\beta}$, the log posterior is

$$\begin{aligned} \ln p(N, \boldsymbol{\beta}, \mathcal{X}_{mis}|\mathcal{Y}, \mathcal{X}_{obs}) \propto & -\frac{1}{2}\sum_{j=1}^{J}(b-\boldsymbol{\beta_j})^T B^{-1}(b-\boldsymbol{\beta_j}) \\ & + \sum_{i=1}^{N}\sum_{j=1}^{J} y_{ij}\ln(\lambda_{ij}) + (1-y_{ij})\ln(1-\lambda_{ij}) \\ & + \ln(1 - \prod_{j=1}^{J}(1-\lambda_{ij})) \end{aligned} \tag{28}$$

Taking the first derivative with respect to $\boldsymbol{\beta_j}$ gives the gradient

$$\frac{\Delta \ln p(N, \boldsymbol{\beta}, \mathcal{X}_{mis}|\mathcal{Y}, \mathcal{X}_{obs})}{\Delta \boldsymbol{\beta_j}} = B^{-1}(b - \boldsymbol{\beta_j}) + \sum_{i=1}^{n} \left(y_{ij} - \lambda_{ij} + \frac{\lambda_{ij} \prod_{j=1}^{J}(1 - \lambda_{ij})}{1 - \prod_{j=1}^{J}(1 - \lambda_{ij})} \right) \boldsymbol{x_i}^T. \quad (29)$$

Notice in the equation above that only the observed data appears. This convenience occurs due to the fact that $\mathrm{E}[y_{ij}|\boldsymbol{y_i} = \mathbf{0}] = 0$ for all individuals (Alho, 1990). Using the gradient, apply the gradient ascent algorithm until convergence to obtain $\boldsymbol{\beta}_{MAP}$.

9 Appendix B: Dirichlet Process of Mixture Normals

The generative process for each observation's covariate is,

$$\boldsymbol{x_i}|z_i \overset{ind}{\sim} \text{MVNormal}(\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \text{ for } i = 1, ..., N \quad (30)$$

$$z_i \overset{iid}{\sim} \text{Discrete}(\{1, 2, ...\}, (\pi_1, \pi_2, ...)) \text{ for } i = 1, ..., N. \quad (31)$$

The probability density function of the covariates can be formally written as,

$$g(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{N} \prod_{k=1}^{\infty} \pi_k \text{ MVNormal}(\boldsymbol{x_i}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \quad (32)$$

In order to have a fully Bayesian approach, we assign priors to the unknown parameters.

$$\boldsymbol{\Sigma_k} \sim \text{InvWishart}(\nu_0, \boldsymbol{\Lambda_0^{-1}}) \tag{33}$$

$$\boldsymbol{\mu_k}|\boldsymbol{\Sigma_k} \sim \text{MVNormal}(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_k}/\kappa_0) \tag{34}$$

$$(\pi_1, \pi_2, ...) \sim SB(\alpha) \tag{35}$$

$$\alpha \sim \text{Gamma}(a_\alpha, b_\beta), \tag{36}$$

where SB(α) is the stick breaking process (Dunson and Xing, 2009). While the problem is defined as having an infinite number of mixture components, we solve the finite-dimension problem with the number of mixtures truncated at an upper bound, K^*. Because the stick breaking process orders the number of mixtures, as long as K^* is set sufficently large, this construction approximates the infinite-dimensional problem. The stick breaking prior can therefore be defined as $\pi_k = V_k \prod_{l<k}(1 - V_l)$, where $V_1, ..., V_{K^*-1} \sim \text{Beta}(1, \alpha)$ and $V_{K^*} = 1$.

Using this generative scheme adds five additional sampling stages to the algorithm presented in section 3.

1) Sample z_i for $i = 1, ..., N$. The latent class label takes integer values from $1, ..., K^*$. To compute the probability of each latent class label for each i,

$$P(z_i = k) = \frac{\pi_k \ \text{MVNormal}(\boldsymbol{x_i}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})}{\sum_{l=1}^{K^*} \pi_l \ \text{MVNormal}(\boldsymbol{x_i}|\boldsymbol{\mu_l}, \boldsymbol{\Sigma_l})}. \tag{37}$$

2) Sample $\boldsymbol{\Sigma_k}$ for $k = 1, ..., K^*$. Define $N_k = \sum_{i=1} 1_{z_i=k}$ which is a count of the number of individuals in the population belonging to latent class, k. Also, define the sufficient statistics, $\bar{\boldsymbol{x}}_{\boldsymbol{k}} = \frac{1}{N_k} \sum_{i=1}^{N_k} \boldsymbol{x_{ik}}$ and $\boldsymbol{S_k} = \sum_{i=1}^{N_k} (\boldsymbol{x_{ik}} - \bar{\boldsymbol{x}}_{\boldsymbol{k}})(\boldsymbol{x_{ik}} - \bar{\boldsymbol{x}}_{\boldsymbol{k}})^T$. Then,

$$\boldsymbol{\Sigma_k} \sim \text{InvWishart}(\nu_{N_k}, \boldsymbol{\Lambda_{N_k}^{-1}}), \tag{38}$$

where $\nu_{N_k} = \nu_0 + N_k$ and $\boldsymbol{\Lambda_{N_k}} = \boldsymbol{\Lambda_0} + \boldsymbol{S_k} + \frac{\kappa_0 N_k}{\kappa_0 + N_k} (\bar{\boldsymbol{x}}_{\boldsymbol{k}} - \boldsymbol{\mu_0})(\bar{\boldsymbol{x}}_{\boldsymbol{k}} - \boldsymbol{\mu_0})^T$.

3) Sample $\boldsymbol{\mu_k}|\boldsymbol{\Sigma_k}$ for $k = 1, ..., K^*$. Using the same defined terms in the previous step,

$$\boldsymbol{\mu_k}|\boldsymbol{\Sigma_k} \sim \text{MVNormal}(\boldsymbol{\mu_{N_k}}, \boldsymbol{\Sigma_k}/\kappa_{N_k}), \tag{39}$$

where $\boldsymbol{\mu_{N_k}} = \frac{\kappa_0}{\kappa_0+N_k}\boldsymbol{\mu_0} + \frac{N_k}{\kappa_0+N_k}\bar{\boldsymbol{x}}_{\boldsymbol{k}}$ and $\kappa_{N_k} = \kappa_0 + N_k$.

4) Sample $(\pi_1, \pi_2, ..., \pi_{K^*})$ for $k = 1, ..., K^*$. Begin by drawing a sample from each of the stochastic components,

$$V_k \sim \text{ Beta}\left(1 + N_k, \alpha + \sum_{l=k+1}^{K} N_l\right) \text{ for } k = 1, ..., K^* - 1. \tag{40}$$

Set $V_{K*} = 1$. Then, $\pi_k = V_k \prod_{l<k}(1 - V_l)$ for all $k = 1, ..., K^*$.

5) Sample α.

$$\alpha \sim \text{ Gamma}\left(a_\alpha + K^* - 1, b_\alpha - \ln(\pi_{K^*})\right). \tag{41}$$

10 Appendix C: LCMCR Model

This section summarizes the Bayesian Non-Parametric Latent-Class Capture-Recapture (LCMCR) derivation and explanations provided in Manrique-Vallier (2016). The LCMCR model is a framework for the capture-recapture (CR) problem that uses the Bayesian nonparametric latent-class model (NPLCM) proposed in Dunson and Xing (2009) to model $f(\boldsymbol{y_i}|\theta)$.

In order to account for unobserved heterogeneity, each individual, i, is modeled such that they belong to a hidden, latent class, z_i, with probability π_k. After determining a latent class, an individual is captured according to a Bernoulli distribution on list j with probability, λ_{jk}. This is known as the latent-class model (Goodman, 1974), and yields the probability mass function

$$f(\boldsymbol{y_i}|\boldsymbol{\lambda}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \lambda_{jk}^{y_{ij}} (1 - \lambda_{jk})^{1-y_{ij}}, \tag{42}$$

where $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)$ and $\boldsymbol{\lambda} = (\lambda_{jk})$ for $j = 1, ..., J$ and $k = 1, ..., K$.

Inserting Equation 42 into Equation 1 yields the following likelihood equation

$$p(\mathcal{Y}|\boldsymbol{\lambda}, \boldsymbol{\pi}, N) = \binom{N}{n} \left[\sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} (1 - \lambda_{jk}) \right]^{N-n} \times \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \lambda_{jk}^{y_{ij}} (1 - \lambda_{jk})^{1-y_{ij}}. \tag{43}$$

The number of latent classes is endogenized with the probability of belonging to each latent class, π_k, being drawn from a "stick-breaking" process (Sethuraman, 1991). The parameter, α, controls the amount of concentration of the probability mass. In other words, larger values of α will lead to a larger number of relevant latent classes. A Beta(1,1) prior distribution is placed on each of the $J \times K^*$ probabilities, and a Gamma(a, b) prior is placed on α.

The model can be summarized through the following hierarchical generative process

$$\begin{aligned}
y_{ij}|z_i &\sim \text{Bernoulli}(\lambda_{jz}) \quad \text{for } j = 1, ..., J \text{ and } i = 1, ..., N \\
z_i &\sim \text{Discrete}(\{1, 2, ...\}, (\pi_1, \pi_2, ...)) \quad \text{for } i = 1, ..., N \\
\lambda_{jk} &\sim \text{Beta}(1, 1) \quad \text{for } j = 1, ..., J \text{ and } k = 1, 2, ... \\
(\pi_1, \pi_2, ...) &\sim \text{SB}(\alpha) \\
\alpha &\sim \text{Gamma}(a, b).
\end{aligned} \tag{44}$$

11 References

Agresti, A. (1994). Simple Capture-Recapture Models Permitting Unequal Catchability and Variable Sampling Effort. *Biometrics*, 50(2):494–500.

Alho, J. M. (1990). Logistic Regression in Capture-Recapture Models. *Biometrics*, 46(3):623.

Alpizar-Jara, R. and Pollock, K. H. (1996). A combination line transect and capture-recapture sampling model for multiple observers in aerial surveys. *Environmental and Ecological Statistics*, 3(4):311–327.

Baker, S. G. (1990). A Simple EM Algorithm for Capture-Recapture Data with Categorical Covariates. *Biometrics*, 46(4):1193.

Ball, P., Asher, J., Sulmont, D., and Manrique, D. (2003). An estimate of the total number of victims killed or disappeared in the armed internal conflict between 1980 and 2000. *AAAS. Report to the Peruvian Truth and Reconciliation Commission (CVR).*

Ball, P., Betts, W., Scheuren, F., Dudukovich, J., and Asher, J. (2002). Killings and Refugee Flow in Kosovo, March - June 1999: A Report to the International Criminal Tribunal for the Former Yugoslavia. *American Association for the Advancement of Science (AAAS).*

Ball, P. and Price, M. (2018). The Statistics of Genocide. *CHANCE*, 31(1):38–45.

Ball, P., Spirer, H. F., and Spirer, L. (2000). *Making the case: investigating large scale human rights violations using information systems and data analysis.* American Association for the Advancement of Science, Washington, D.C.

Basu, S. (2001). Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence. *Biometrika*, 88(1):269–279.

Bonner, S. and Schofield, M. (2014). MC(MC)MC: exploring Monte Carlo integration within MCMC for mark-recapture models with individual covariates. *Methods in Ecology and Evolution*, 5(12):1305–1315.

Borchers, D. L., Zucchini, W., and Fewster, R. M. (1998). Mark-Recapture Models for Line Transect Surveys. *Biometrics*, 54(4):1207–1220.

Castledine, B. J. (1981). A Bayesian Analysis of Multiple-Recapture Sampling for a Closed Population. 68:197–210.

Chapman, D. G. (1954). The Estimation of Biological Populations. *The Annals of Mathematical Statistics*, 25(1):1–15.

Darroch, J. N. (1958). The Multiple-Recapture Census: I. Estimation of a Closed Population. *Biometrika*, 45(3/4):343–359.

Darroch, J. N., Fienberg, S. E., Glonek, G. F. V., and Junker, B. W. (1993). A Three-Sample Multiple-Recapture Approach to Census Population Estimation With Heterogeneous · catchability. *Journal of the American Statistical Association*, page 13.

Dunson, D. B. and Park, J.-H. (2008). Kernel stick-breaking processes. *Biometrika*, 95(2):307–323.

Dunson, D. B. and Xing, C. (2009). Nonparametric Bayes Modeling of Multivariate Categorical Data. *Journal of the American Statistical Association*, 104(487):1042–1051.

Fienberg, S. E. (1972). The Multiple Recapture Census for Closed Populations and Incomplete 2k Contingency Tables. *Biometrika*, 59(3):591–603.

Fienberg, S. E., Johnson, M. S., and Junker, B. W. (1999). Classical Multilevel and Bayesian Approaches to Population Size Estimation Using Multiple Lists. *Journal of Royal Statistical Society*, 162(3):383–405.

Geiger, H. and Werner, A. (1924). Die Zahl der von Radium ausgesandten a-Teilchen. *Zeitschrift für Physik*, 21:187–201.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*. Chapman & Hall/CRC, third edition.

George, E. I. (1992). Capture-Recapture Estimation Via Gibbs Sampling. *Biometrika*, 79(4):677–683.

Goldberg, J. D. and Wittes, J. T. (1978). The Estimation of False Negatives in Medical Screening. *Biometrics*, 34(1):77–86.

Goodman, L. A. (1974). Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models. *Biometrika*, 61(2):215–231.

Guberek, T., Guzman, D., Price, M., Lum, K., and Ball, P. (2010). To Count the Uncounted: An Estimation of Lethal Violence in Casanare. *Benetech Human Rights Program*, pages 1–31.

Henderson, P. A. and Southwood, T. (2016). *Ecological Methods*. John Wiley and Sons, 4 edition.

Herliansyah, R., King, R., and King, S. (2022). Laplace Approximations for Capture–Recapture Models in the Presence of Individual Heterogeneity. *Journal of Agricultural, Biological and Environmental Statistics*, 27(3):401–418.

Horvitz, D. G. and Thompson, D. J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, 47(260):663–685.

Huggins, R. M. (1989). On the Statistical Analysis of Capture Experiments. *Biometrika*, 76(1):133–140.

Ishwaran, H. and James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96(453):161–173.

Jeffreys, H. (1967). *Theory of Probability*. Clarendon Press, Oxford, 3rd edition.

Khabsa, M. and Giles, C. L. (2014). The Number of Scholarly Documents on the Public Web. *PLoS ONE*, 9(5):e93949.

King, R. and Brooks, S. P. (2008). On the Bayesian Estimation of a Closed Population Size in the Presence of Heterogeneity and Model Uncertainty. *Biometrics*, 64(3):816–824.

King, R., McClintock, B. T., Kidney, D., and Borchers, D. (2016). Capture–recapture abundance estimation using a semi-complete data likelihood approach. *The Annals of Applied Statistics*, 10(1):264–285.

Lawrence, S. and Giles, C. L. (1998). Searching the World Wide Web. *Science*, 280(5360):98–100.

Lincoln, F. C. (1930). Calculating Waterfowl Abundance on the Basis of Banding Returns. *United States Department of Agriculture*, Circular 118.

Manrique-Vallier, D., Ball, P., and Sadinle, M. (2020). Capture-Recapture for Casualty Estimation and Beyond: Recent Advances and Research Directions.

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.

Manrique-Vallier, D. and Fienberg, S. E. (2008). Population Size Estimation Using Individual Level Mixture Models. *Biometrical Journal*, 50(6):1051–1063.

Manrique-Vallier, D. (2016). Bayesian Population Size Estimation using Dirichlet Process Mixtures. *Biometrics*, 72(4):1246–1254.

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*, (62,):3–135.

Petersen, C. J. (1895). The yearly immigration of young plaice into the limfjord from the german sea. *The Danish Biological Station*.

Pledger, S. (2000). Unified Maximum Likelihood Estimates for Closed Capture-Recapture Models Using Mixtures. *Biometrics*, 56(2):434–442.

Pollock, K. H. (2002). The use of auxiliary variables in capture-recapture modelling: An overview. *Journal of Applied Statistics*, 29(1-4):85–102.

Pollock, K. H., Hines, J. E., and Nichols, J. D. (1984). The Use of Auxiliary Variables in Capture-Recapture and Removal Experiments. *Biometrics*, 40(2):329–340.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Polya-Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

Roberts, H. V. (1967). Informative Stopping Rules and Inferences about Population Size. *Journal of the American Statistical Association*, 62:763–775.

Royle, J. A. (2009). Analysis of Capture-Recapture Models with Individual Covariates Using Data Augmentation. *Biometrics*, 65(1):267–274.

Royle, J. A., Dorazio, R. M., and Link, W. A. (2007). Analysis of Multinomial Models With Unknown Index Using Data Augmentation. *Journal of Computational and Graphical Statistics*, 16(1):67–85.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63(3):581–592.

Sanathanan, L. (1972). Models and Estimation Methods in Visual Scanning Experiments. *Technometrics*, 14(4):813–829.

Sandland, R. L. and Cormack, R. M. (1984). Statistical Inference for Poisson and Multinomial Models for Capture- Recapture Experiments. *Biometrika*, 71(1):27–33.

Schaefer, M. B. (1951). Estimation of Size of Animal Populations by Marking Experiments. *Fishery Bulletin*, 52(69):191–203.

Schnabel, Z. E. (1938). The Estimation of Total Fish Population of a Lake. *The American Mathematical Monthly*, 45(6):348.

Sekar, C. C. and Deming, E. (1949). On a Method of Estimating Birth and Death Rates and the Extent of Registration. *Journal of the American Statistical Association*, 44(245):101–115.

Sethuraman, J. (1991). A Constructive Definition of Dirichlet Priors:. Technical report, Defense Technical Information Center, Fort Belvoir, VA.

Smith, P. J. (1991). Bayesian Analyses for a Multiple Capture-Recapture Model. *Biometrika*, 78(2):399–407.

Stoklosa, J., Hwang, W.-H., Wu, S.-H., and Huggins, R. (2011). Heterogeneous Capture-Recapture Models with Covariates: A Partial Likelihood Approach for Closed Populations. *Biometrics*, 67(4):1659–1665.

Tilling, K. and Sterne, J. A. C. (1999). Capture-Recapture Models Including Covariate Effects. *American Journal of Epidemiology*, 149(4):392–400.

Woodbury, M., Clive, J., and Garson, A. (1978). Mathematical typology: a grade of membership technique for obtaining disease definition. *Computers and Biomedical Research*, 11(3):277–298.

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(5).

Zwane, E. and van der Heijden, P. (2003). Implementing the parametric bootstrap in capture–recapture models with continuous covariates. *Statistics & Probability Letters*, 65(2):121–125.

Zwane, E. and van der Heijden, P. (2005). Population estimation using the multiple system estimator in the presence of continuous covariates. *Statistical Modelling*, 5(1):39–52.

Curriculum Vitae