Population Sized Graphical Record Linkage

Andee Kaplan and Rebecca C. Steorts

April 12, 2018

Abstract

1. We propose a generalized framework for entity resolution.
2. Connection with Gibbs partition models and that of Steorts (2015).
3. We make the first mention of the label switching issue to our knowledge of record linkage, highlighting key ways of solving this problem in practice.
4. We also make the first construction of solve the down stream task.
5. The proposed methodology is applied on a synthetic data set, where comparisons are made to supervised learning methods.
6. Can any connection be made to the degradation that AK has been working on for blocking? I think perhaps yes if the Bayesian method is always beating the others since then it should always be preferred.

1 Introduction

Very often information about social entities is scattered across multiple databases. Combining that information into one database can result in enormous benefits for analysis, resulting in richer and more reliable conclusions. In most practical applications, however, analysts cannot simply link records across databases based on unique identifiers because they are not a part of some databases or are not available due to privacy concerns. In such cases, analysts often use *record linkage* (*entity resolution* or *de-duplication*) — the process of linking records corresponding to unique entities either within a single database or across multiple data sources. Record linkage is not only a crucial task for social science and industrial applications, but is a challenging statistical and computational problem itself, because many databases contain errors (noise, lies, omissions, duplications, etc.), and the number of parameters to be estimated grows with the number of records [? ? ? ? ? ? ? ? ? ? ?]. In addition, in many cases, record linkage is just the first step to underlying questions posed by data analysts, and the second component requires the integration of additional post-linkage analyses, where the goal is to estimate a population parameter.

Many record linkage methods are an extension of the Fellegi-Sunter (FS) approach, which computes pairwise probabilities of matching for all pairs of records using a likelihood ratio test [? ?]. While modern FS methods are used today, such implementations assume that only two databases can be linked and that there are no duplicates within each database [? ? ?]. In addition, such approaches are sensitive to the choice of the threshold, are not easily generalizable to a large class of models, and the record linkage uncertainty does not easily propagate into subsequent analyses.

Bayesian methods have been utilized in record linkage due to their flexibility and exact error propagation; however, they have been limited primarily to two-database matching, due to scalability issues and model misspecification [? ? ? ? ?]. These contributions, while valuable, do not easily generalize to multiple databases and to duplication within databases. Specifically, [? ?] developed a fully hierarchical-Bayesian approach to record linkage using Dirichlet prior distributions over latent attributes and assuming a data distortion model. The attributes of the latent entities, the number of latent entities, the edges linking records to latents, etc., all have posterior distributions, and it is easy to sample from these distributions for uncertainty quantification or error propagation. More recently, [?] extended their approach to both categorical and text data using an empirically motivated prior (blink), which beat many supervised methods (e.g., random forests, Bayesian Adaptive Regression Trees, logistic regression) in terms of accuracy when the training data is 10 percent (or less) of the total amount of data.

The only method to our knowledge that does a fully Bayesian approach of record linkage and capture recapture is that of [**?**], where the authors apply this to the scenario of two data sets and continuous data. There are many advantages of such a method and we combine the best of both worlds here by proposing a general Bayesian record linkage framework in conjunction with a general post-linkage framework, and specifically for a capture-recapture problem. AK:Could you add a background of capture-recapture here. It's okay if it's too long. I can shorten it.

In this paper, we highlight a general class of semi-parametric Bayesian graphical record linkage models for use within a generalized framework for post-linkage analysis. Specifically, we discuss how entities can be uniquely identified from such graphical models, the error from the record linkage process can be propagated exactly, and posterior linkage probabilities can inform population estimation through use of a capture-recapture model. Finally, we illustrate our approach on a set of simulated data sets, as well as one real data set, providing comparisons in the literature, as well as future directions for the work.

2 Generalized Bayesian Graphical Record Linkage

In this section, we specify a generalized Bayesian graphical model for record linkage, and then introduce the empirically motivated models of [**?**].

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ represent records comprised of D databases, indexed by i. The ith database has n_i observed records, indexed by j. Each record corresponds to one of M latent entities, indexed by j'. Each record or latent entity has values on p fields, indexed by ℓ, and are assumed to be categorical or string and have the same fields across all records and entities [**? ?**]. Let M_ℓ denote the number of possible categorical values for the ℓth field. Let $X_{ij\ell}$ denote the observed value of the ℓth field for the jth record in the ith database, and $Y_{j'\ell}$ denotes the true value of the ℓth field for the j'th latent entity. Then Λ_{ij} denotes the latent entity to which the jth record in the ith database corresponds, i.e., $X_{ij\ell}$ and $Y_{j'\ell}$ represent the same entity if and only if $\Lambda_{ij} = j'$. Then $\boldsymbol{\Lambda} = \{\Lambda_{ij} : i = 1, \ldots, D, j = 1, \ldots, n_i\}$ denotes the Λ_{ij} collectively. Distortion is denoted by $z_{ij\ell} = I(X_{ij\ell} \neq Y_{\Lambda_{ij}\ell})$, where $I(\cdot)$ represents the indicator function (e.g., $I(X_{ij\ell} = m)$ is 1 when the ℓth field in record j in file i has the value m), and let δ_a denote the distribution of a point mass at a (e.g., $\delta_{y_{\Lambda_{ij}\ell}}$).

Using the aforementioned notation, we assume the following model:

$$\begin{aligned}
X_{ij\ell} \mid \Lambda_{ij},\, Y_{\Lambda_{ij}\ell},\, z_{ij\ell} &\overset{\text{ind}}{\sim} F \\
Y_{j'\ell} &\overset{\text{ind}}{\sim} H \\
z_{ij\ell} \mid \beta_{i\ell} &\overset{\text{ind}}{\sim} \text{Bernoulli}(\beta_{i\ell}) \\
\beta_{i\ell} &\overset{\text{ind}}{\sim} \text{Beta}(a, b) \\
\Lambda_{ij} &\overset{\text{ind}}{\sim} \text{KP}, \qquad (1)
\end{aligned}$$

where F and H are generic distributions, all distributions are independent of each other, and a, b are assumed known. The marginal prior for $\mathbf{\Lambda}$ is a Kolchin partition (KP) model, which is a general class of BNP models [**?**], and is closely related to Gibbs-type partitions ([**?** , theorem 1.2]). Within this model class, the number of unique latent entities, M, has a prior placed on it and the number of records clustered to each individual is directly modeled with another prior. (Observe that record linkage and de-duplication are both simply a question of whether $\Lambda_{i_1,j_1} = \Lambda_{i_2,j_2}$, where $i_1 \neq i_2$ for record linkage and $i_1 = i_2$ for de-duplication.) [**?**] showed that such models exhibit the microclustering property, meaning that the size of the largest cluster grows sub-linearly with the total number of records.

2.1 The Empirically Motivated Prior

Given the noisy text and categorical data associated with record linkage, one fully Bayesian framework for model (**??**) if that of [**?**] due to its flexibility to handle such data and superiority over semi-supervised methods. [**?**] assumes fields $1, \ldots, p_s$ are string-valued, while fields $p_s + 1, \ldots, p_s + p_c$ are categorical, where $p_s + p_c = p$ is the total number of fields. They assume an empirical Bayesian distribution on the latent field values, as well as for the categorical record field values. For each $\ell \in \{1, \ldots, p_s + p_c\}$, let S_ℓ denote the set of *all* values for the ℓth field that occur anywhere in the data, i.e., $S_\ell = \{X_{ij\ell} : 1 \leq i \leq D, 1 \leq j \leq n_i\}$, and let $\alpha_\ell(w)$ equal the empirical frequency of value w in field ℓ. Let G_ℓ denote the empirical distribution of the data in the ℓth field from all records in all databases combined. So, if a random variable W has distribution G_ℓ, then for every $w \in S_\ell$, $P(W = w) = \alpha_\ell(w)$. Hence, let G_ℓ be the prior for each latent entity $Y_{j'\ell}$. For string-valued record fields $\ell = 1, \ldots, p_s$ a distortion process is added, which is captured in the prior probability as

$$P(X_{ij\ell} = w \mid \Lambda_{ij}, Y_{\Lambda_{ij}\ell}, z_{ij\ell}) = \frac{\alpha_\ell(w)\, \exp[-c\, d(w, Y_{\Lambda_{ij}\ell})]}{\sum_{w \in S_\ell} \alpha_\ell(w)\, \exp[-c\, d(w, Y_{\Lambda_{ij}\ell})]},$$

where $c > 0$ is a fixed normalizing constant corresponding to an arbitary distance metric $d(\cdot, \cdot)$. Denote this distribution by $F_\ell(Y_{\Lambda_{ij}\ell})$. The full hierarchical model for record linkage becomes

$$\begin{aligned} X_{ij\ell} \mid \Lambda_{ij},\, Y_{\Lambda_{ij}\ell},\, z_{ij\ell} &\overset{\text{ind}}{\sim} \begin{cases} \delta(Y_{\Lambda_{ij}\ell}) & \text{if } z_{ij\ell} = 0 \\ F_\ell(Y_{\Lambda_{ij}\ell}) & \text{if } z_{ij\ell} = 1, \ell \leq p_s \\ G_\ell & \text{if } z_{ij\ell} = 1, \ell > p_s \end{cases} \\ Y_{j'\ell} &\overset{\text{ind}}{\sim} G_\ell \\ z_{ij\ell} \mid \beta_{i\ell} &\overset{\text{ind}}{\sim} \text{Bernoulli}(\beta_{i\ell}) \\ \beta_{i\ell} &\overset{\text{ind}}{\sim} \text{Beta}(a, b) \\ \Lambda_{ij} &\overset{\text{ind}}{\sim} \text{Uniform}(1, \ldots, M), \end{aligned} \tag{2}$$

where all distributions are also independent of each other and a, b, M are assumed known hyperparameters. In the case of M, we assume $M = \sum_{i=1}^{D} n_i$.

2.2 Equivalance of Partition Models

It is not obvious that Model (**??**) is an example of Model (**??**). Specifically, it is not immediately clear that the iid Uniform prior on Λ_{ij} falls within the class of KP priors. In this section, we show their equivalence.

Theorem 1. *There exists a KP model prior that corresponds to the iid Uniform prior on Λ_{ij} as specified in Model (**??**).*

Proof. Let $N_1, \ldots, N_M$ denote the number of individual records clustered to each latent individual. Recall, there are M latent individuals. We can specify the following KP model,

$$M \sim \text{Poisson}(\alpha) \tag{3}$$

$$N_1, \ldots, N_M \overset{iid}{\sim} \text{Poisson}(\gamma) \tag{4}$$

where the cluster assignments $\boldsymbol{\Lambda} = \{\lambda_{ij} : i = 1, \ldots, D, j = 1, \ldots, n_i\}$ are drawn uniformly at random from the permutations of

$$(\underbrace{1, \ldots, 1}_{N_1 \text{ times}}, \underbrace{2, \ldots, 2}_{N_2 \text{ times}}, \ldots, \underbrace{M, \ldots, M}_{N_M \text{ times}}). \tag{5}$$

This is a limiting case of the NBNB model specified in [?], and is itself a KP prior. Note that the cluster assignments of the KP model, $\mathbf{\Lambda}$, correspond to the latent entities in Model (**??**) and within the KP paradigm, $n = \sum_{m=1}^{M} N_m$ is a random variable, even though in Model (**??**) it corresponds to the total number of records in the dataset.

Our goal is to show that the distribution of $\mathbf{\Lambda} \mid n, M$ resulting from the KP model (**??**) corresponds to the iid Uniform$(1, \ldots, M)$ from Model (**??**), i.e. $P(\mathbf{\Lambda} \mid n, M) = M^{-n}$.

To see this, first note that $P(\mathbf{\Lambda} \mid n, M) = \frac{P(\mathbf{\Lambda}, n|M)}{P(n|M)}$ due to the definition of conditional probability. Additionally, due to the selection of the cluster assignments $\mathbf{\Lambda}$ through uniformly at random selection from the permutations of the vector (**??**),

$$P(\mathbf{\Lambda} \mid N_1, \ldots, N_M, M) = \frac{\prod\limits_{m=1}^{M} N_m!}{(\sum\limits_{m=1}^{M} N_m)!} = \frac{\prod\limits_{m=1}^{M} N_m!}{n!}.$$

Then,

$$\begin{aligned}
P(\mathbf{\Lambda}, n \mid M) &= P(\mathbf{\Lambda}, n, N_1, \ldots, N_M \mid M) && \text{(Duplicate information)} \\
&= P(\mathbf{\Lambda}, N_1, \ldots, N_M \mid M) && \text{(Duplicate information)} \\
&= P(\mathbf{\Lambda} \mid N_1, \ldots, N_M, M) P(N_1, \ldots, N_M \mid M) \\
&= \frac{\prod\limits_{m=1}^{M} N_m!}{n!} \prod_{m=1}^{M} \frac{1}{N_m!} \gamma^{N_m} e^{-\gamma} \\
&= \frac{1}{n!} \gamma^n e^{-M\gamma} \\
&= M^{-n} P(n|M),
\end{aligned}$$

where the last line holds because $n = \sum_{m=1}^{M} N_m$ is the sum of conditionally iid Poisson(γ) distributions.

Thus, $P(\mathbf{\Lambda} \mid n, M) = M^{-n}$, and equivalence holds between the models. □

Remark 1. *It should be noted that the proof of Theorem* **??** *is independent of the distribution placed on M in the KP model, and so the result will hold regardless of this distributional choice.*

3 The Label Switching Issue

Talk about the label switching issue, which is avoided by looking at summaries of the linkage structure or by the MPMMS. Talk about how this generalizes to all partition models. Write out the details regarding the MPMMS here formally for generalized partition models.

3.1 The Downstream Task

It seems that we can look at the downstream task quite easily by using the MPMMS and having these point to the latent entity. This would be a start. If time permits, we could use the prototype method to see how this compares.

4 Experiments

We could do an experimental section identifying the individuals with posterior probabilities and making a list of these. We could compare this to supervised methods, where nothing is available. We could leave the next part for the follow up paper.

5 Generalized Post-linkage Analysis

Typically one is interested in record linkage as a pre-processing tool, and other post-linkage tasks are performed afterwards, e.g., linear regression, capture-recapture, or other types of statistical analyses, where one wishes to estimate a parameter about a population $\boldsymbol{\eta}$. It is of great importance to assess the record linkage uncertainty and propagate this error into the post-linkage task.

It is natural to quantify the uncertainty of the record linkage process with regards to $p(\boldsymbol{\Lambda} \mid \boldsymbol{X})$, and without loss of generality we assume this posterior is a proper distribution. Then, assuming we know the true value of the linkage structure $\boldsymbol{\Lambda}$, let us represent our beliefs about the population parameter $\boldsymbol{\eta}$ by $p_C(\boldsymbol{\eta} \mid f(\boldsymbol{\Lambda}))$, which is the posterior obtained from a post-linkage model C with some likelihood and prior distribution given $f(\boldsymbol{\Lambda})$ a function of the linkage structure. In this setup, we can quantify the record linkage uncertainty in the population parameter by the following quantity:

$$U(\boldsymbol{\eta}) =: E_{\boldsymbol{\Lambda}|\boldsymbol{X}}[p_C(\boldsymbol{\eta} \mid f(\boldsymbol{\Lambda}))] = \sum_{\boldsymbol{\Lambda}} p_C(\boldsymbol{\eta} \mid f(\boldsymbol{\Lambda}))p(\boldsymbol{\Lambda} \mid \boldsymbol{X}), \tag{6}$$

where f is some function of $\boldsymbol{\Lambda}$. Namely, $U(\boldsymbol{\eta})$ is the marginal posterior distribution of $\boldsymbol{\eta}$ assuming a linkage model and a joint prior $p(\boldsymbol{\eta} \mid f(\boldsymbol{\Lambda})p(\boldsymbol{\Lambda})$. Intuitively, we can view $U(\boldsymbol{\eta})$ as the expected posterior distribution of the population parameter of interest, where we have averaged over the posterior of the linkage structure.

Next, the total variability denoted by $U(\boldsymbol{\eta})$ can also be easily decomposed. Recall $U(\boldsymbol{\eta}) = p(\boldsymbol{\eta} \mid \boldsymbol{X})$. Then

$$\mathrm{Var}(\boldsymbol{\eta} \mid X) = \mathrm{Var}_{\boldsymbol{\Lambda}|\boldsymbol{X}}[E[\boldsymbol{\eta} \mid \boldsymbol{\Lambda}]] + E_{\boldsymbol{\Lambda}|\boldsymbol{X}}[\mathrm{Var}[\boldsymbol{\eta} \mid \boldsymbol{\Lambda}]], \tag{7}$$

where $\mathrm{Var}_{\boldsymbol{\Lambda}|\boldsymbol{X}}[E[\boldsymbol{\eta} \mid \boldsymbol{\Lambda}]]$ corresponds to the record linkage uncertainty due to population size estimation and $E_{\boldsymbol{\Lambda}|\boldsymbol{X}}[\mathrm{Var}[\boldsymbol{\eta} \mid \boldsymbol{\Lambda}]]$ corresponds to the variability associated from a Bayesian method due to estimating $\boldsymbol{\eta}$. In practice, $U(N)$ and both quantities in (**??**) must be estimated by Markov chain Monte carlo.

6 Capture-recapture

Capture-recapture models have been developed to estimate population size from multiple samples taken from a single, closed population. We will perform capture-recapture post-linkage with the goal of estimating the population size in a Bayesian framework that takes the uncertainty of linkage into account. First, we will introduce some additional notation for this post-linkage analysis and give a general model formulation for capture recapture. Section **??** gives details on two specific model formulations that we will employ within the data experiments of Section **??**.

We assume a closed population of size N, and the D samples taken correspond to the D databases that comprise our records $\boldsymbol{X}_n$. In this case, a "capture" simply corresponds to occurence in a database. Let Q be the full $N \times D$ capture matrix, which has one row for each population unit $i = 1, \ldots, N$ and one column for each database $j = 1, \ldots, D$. Each entry in the capture matrix q_{ij} takes value one if individual i occurs in database j and zero otherwise. The entries of Q are thus $N * D$ Bernoulli trials, $q_{ij} \overset{ind}{\sim} \text{Bern}(p_{ij})$, where p_{ij} is the probability of occurence of individual i in database j. Let $P = \{p_{ij}\}$ for $i = 1, \ldots, N$ and $j = 1, \ldots, D$ denote the matrix of probabilities. Then the ith row of Q is the *capture vector* $\boldsymbol{q}_i$ for individual i and there are 2^D unique possible capture vectors, $\boldsymbol{q}_i \in \{0, 1\}^D$. The zero capture vector $\boldsymbol{q} = \boldsymbol{0}$ indicates that an individual was not captured in any database, and is unobservable. In order to estimate N, we can reframe the problem to estimating $c_{\boldsymbol{0}} = \sum_{i=1}^{n} \boldsymbol{I}(\boldsymbol{q}_i = \boldsymbol{0})$ the number of individuals with zero capture vector since $N = c_{\boldsymbol{0}} + n'$, where recall that n' is the number of unique individuals captured within the D databases corresponding to the number of latent individuals.

Then, conditional on N and P, the full capture matrix Q has likelihood

$$p(Q|P, N) = \frac{N!}{\prod_{\boldsymbol{q}} c_{\boldsymbol{q}}} \prod_{i=1}^{N} \prod_{j=1}^{D} p_{ij}^{q_{ij}} (1 - p_{ij})^{1 - q_{ij}}$$

where $c_{\boldsymbol{q}}$ is the number of individuals with capture vector equal to $\boldsymbol{q}$. Typically the capture matrix P is modelled using some set of parameters, $\boldsymbol{\theta}$, which can depend on i, j, or both. This is captured in the data generation distribution as

$$p(Q|\boldsymbol{\theta}, N) = \frac{N!}{\prod_{\boldsymbol{q}} c_{\boldsymbol{q}}} \prod_{i=1}^{N} f(\boldsymbol{q}_i|\boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ will be of reduced dimension than P for identifiability. The data generation process is the subject of much study in the literature due to its ability to capture dependence or heterogeneity among individuals or between lists (See [**? ? ?**] for reviews). One method that has been proposed for dealing with heterogeneity is the use of mixture models to represent heterogeneous populations that arise from the aggregation of two or more homogeneous populations [**? ?**]. We employ the Bayesian Non-Parametric

Latent Class (NPLCM) model of [?] for heterogeneity, which uses the Dirichlet Process mixture to allow for a data-based selection of the included number of homogeneous populations, or mixture components.

6.1 Independence Model

For comparison to the infinite dimensional mixture model, we first present the independence model for the capture distribution,

$$f(\boldsymbol{q}|\boldsymbol{\lambda}) = \prod_{j=1}^{D} \lambda_j^{q_j}(1-\lambda_j)^{1-q_j}. \tag{8}$$

This model assumes that the database inclusion processes are independent and the probability of inclusion for each individual within a database are identical. When these assumptions are violated, models of this form will produce unreliable estimates [?].

When heterogeneity is present, a common approach is to stratify the population into homogeneous classes where the independence model is expected to hold [?]. Typically covariates are used to acheive this, but in the absense of covariate information, a latent variable approach can be used. The data generating distribution is then given as

$$f(\boldsymbol{q}|\boldsymbol{\lambda},\boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{D} \lambda_{jk}^{q_j}(1-\lambda_{jk})^{1-q_j}, \tag{9}$$

where $\boldsymbol{\lambda} = \{\lambda_{jk} : j = 1,\ldots,D, k = 1,\ldots,K\}$ with $\lambda_{jk} \in (0,1)$, $\boldsymbol{\pi} = \{\pi_k : k = 1,\ldots,K\}$ with $\sum_{k=1}^{K}\pi_k = 1$ and $\pi_k > 0$ are the strata probabilities. This is a finite mixture model with each component as an independent model. When K is properly selected, this model can perform well [?]. Using the method of [?] (Section ??), we can avoid having to select K explicitly.

6.2 Modeling Heterogeneity using the NPLCM

The NPLCM is an extension of model (??) proposed by [?] that uses an infite mixture model to avoid the need to a priori specify K, while enforcing sparsity of the components by concentrating most of the probability mass onto a small finite set of latent classes (through use of a Dirichlet process prior). In practice, a finite-dimensional approximation is used, where a large-enough upper bound for the number of classes, K^*, is specified. Sensitivity to this upper bound was assessed in the context of capture-recapture in [?], and as long as K^* is large enough, there is no noticable impact on the estimates. The NPLCM model for capture-recapture is obtained by combining model (**??**) with a Dirichlet process prior for the latent classes

$$f(\boldsymbol{q}|\boldsymbol{\gamma}, \boldsymbol{\pi}) = \sum_{k=1}^{K^*} \pi_k \prod_{j=1}^{D} \gamma_{jk}^{q_j} (1 - \gamma jk)^{1-q_j},$$

where $(\pi_1, \dots, \pi_{K^*}) \sim \text{SB}_{K^*}(\alpha)$ and $\text{SB}_{K^*}(\alpha)$ is the finite dimensional approximation to the stick breaking prior with parameter $\alpha > 0$. This approximation is acheived by making $\pi_k = V_k \prod_{h<k}(1 - V_h)$ for $V_{K^*} = 1$ and $V_1, \dots, V_{K^*-1} \overset{iid}{\sim} \text{Beta}(1, \alpha)$.

The remainder of the Bayesian specification for the model follows from [**?**],

$$\begin{aligned} \gamma_{jk} &\overset{iid}{\sim} \text{Beta}(a_\gamma, b_\gamma) \\ \alpha &\sim \text{Gamma}(a_\alpha, b_\alpha) \\ p(N) &\propto \frac{1}{N}, \end{aligned}$$

where $a_\gamma, b_\lambda, a_\gamma, b_\alpha$ are hyperparameters, considered known, and the prior on population size N is the Jeffreys' prior. In accordance with the discussion in [**?**], we let $a_\gamma = 1, b_\gamma = 1, a_\alpha = 0.25, b_\alpha = 0.25$.

6.3 Capture-recapture as Post-linkage Analysis

In the capture-recapure problem framed in this section, the capture vectors $\boldsymbol{q}_i$ are assumed known for each individual. By incorporating capture-recature as a post-linkage analysis, we now need to consider the capture vectors as a function of $\boldsymbol{\Lambda}$ the linkage structure to incorporate the uncertainty from record linkage, via the posterior distribution. This function corresponds to the function of the linkage structure defined in Section **??** and used in (**??**) to quantify the record linkage uncertainty in the population parameter, which in this application corresponds to the population size, N.

$$f(\boldsymbol{\Lambda}) = \{f_{j'}(\boldsymbol{\Lambda})\}_{j'=1,\dots,N},$$

where

$$f_{j'}(\boldsymbol{\Lambda}) = \{I(j' \in \{\lambda_{ij} : j = 1, \dots, n_i\})\}_{i=1,\dots,D}$$

is a binary vector of length D that corresponds to the capture history of individual j'. For $j' = M + 1, \dots, N$, this function will necessarily equal $\mathbf{0}$

because these are unobserved individuals, and so $f^{(M)}(\mathbf{\Lambda})$ the first M rows of $f(\mathbf{\Lambda})$ is the observable quantity, given M. When we combine capture-recapture with record linkage, however, M is a random variable that corresponds to the number of latent individuals captured in all databases and so, we have a posterior distribution of $f^{(M)}(\mathbf{\Lambda})$ given M.

7 Experiments

In this section, we [[insert words here]].

7.1 Simulated Population with Varying Duplication

Talk about the experimental set up regarding the JASA papers. Talk about how we are creating a population. Draw a picture regarding the process if time permits to go along with the description.

For the simulated population, we consider the record linkage method of `blink` and the capture recapture method of `NPLCM`. For a record linkage method, we report the following evaluation metrics: (a) recall, (b) precision, (c) posterior mean, (d) posterior standard deviation, (e) posterior density compared to ground truth, and (f) record linkage error. For a capture recapture method, we report the following evaluation metrics: [[let's chat about these Andee]].

7.2 Results

A Graphical Representation of the Empirically Motivated Record Linkage Model

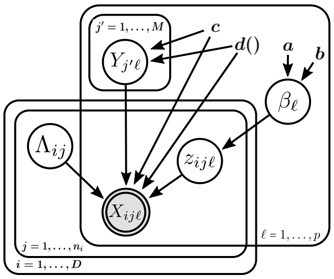

Figure 1: Graphical representation of model (**??**).