BAYESIAN RECORD LINKAGE

RACHEL ANDERSON

1. Introduction

Ask any economist who works with survey or administrative data how they handle imperfect matches when performing a one-to-one merge, and you will get a different answer. Some pick the matches that they believe are most likely to be correct. Some estimate the same model using multiple configurations of matched data. Others avoid the issue entirely, by dropping all matches that are flagged during the automated merge processes in R or Stata. There is no one-size-fits-all solution for data merging, nor a formal theory about how subjective decisions in the merging process may introduce bias or uncertainty into estimation and inference in economic models.

The process of joining records from multiple data sources that describe the same entity appears in many fields, such as statistics, computer science, operations research, and epidemiology, under many names, such as record linkage, data linkage, entity resolution, instance identification, de-duplication, merge/purge processing, and reconciliation. In econometrics, the process is commonly referred to as "data combination", although literature notably lacking, despite the increasing availability of large administrative datasets.

There are now many solutions to the record linkage problem, and yet little understanding of how these techniques may introduce sample selection bias, and how researchers should handle these issues in practice. Examining Bayesian record linkage tools is a promising place to start, as they provide a coherent framework for incorporating uncertainty at every point of the analysis.

In this paper, I review and implement the Bayesian record linkage procedures from Larsen and Rubin (2001) and Sadinle (2017) to obtain a posterior distribution over the bipartite matching/sets of matches vs. non-matches. I use simulated data (and maybe a real data set if I have time). I analyze adjustments for what makes a better match and compare them to traditional techniques.

Then, with these posteriors, I perform a basic regression analysis using methods from Larsen and Lahiri (2005) and Liseo/Tancredi (2015) about how to propagate uncertainty from the matches into a basic regression analysis.

2. Probabilistic Record Linkage via Mixture Models

2.1. **Setup.**

Consider two datafiles X_1 and X_2 that record information from two overlapping sets of individuals or entities. The files originate from two record-generating processes that may induce errors and missing values, so that identifying which individuals are represented in both X_1 and X_2 is nontrivial. I assume that individuals appear at most once in each datafile[1], so that the goal of record linkage is to identify which records in files X_1 and X_2 refer to the same entities.

Suppose that files X_1 and X_2 contain n_1 and n_2 records, respectively, and without loss of generality that $n_1 \geq n_2$. Denote also the number of entities represented in both files as n_M, so that $n_2 \geq n_M \geq 0$.

[1] If not, one would perform first de-duplication, another topic

According to the probabilistic record linkage framework of Fellegi and Sunter (1969), the set of ordered record pairs $X_1 \times X_2$ is the union of two disjoint sets, *matches* (M) and *non-matches* (U):

$$\begin{aligned} M &= \{(i,j) : i \in X_1, j \in X_2, \Delta_{ij} = 1\} \\ U &= \{(i,j) : i \in X_1, j \in X_2, \Delta_{ij} = 0\} \end{aligned}$$

A record pair $(i,j) \in X_1 \times X_2$ is evaluated according to L different comparison criteria, represented by the comparison vector $\gamma_{ij} = (\gamma_{ij}^1, \ldots, \gamma_{ij}^\ell, \ldots, \gamma_{ij}^L)$. The comparison criteria may be binary, as in "i and j have the same birthday", or factor variables that account for partial agreement between strings (see Jaro-Winkler reference here).

The probability of observing a particular configuration of γ_{ij} can be modeled as arising from the mixture distribution:

$$P(\gamma_{ij}) = P(\gamma_{ij}|M)p_M + P(\gamma_{ij}|U)p_U \tag{1}$$

where $P(\gamma_{ij}|M)$ and $P(\gamma_{ij}|U)$ are the probabilities of observing the pattern γ_{ij} conditional on the record pair (i,j) belonging to M or U , respectively. The proportions p_M and $p_U = 1 - p_M$ are the marginal probabilities of observing a matched or unmatched pair in the population of record pairs.

Since membership to M or U is not actually observed, a convenient way of simultaneously estimating p_M, p_U and performing classification is via mixture modeling, with mixture distributions $P(\gamma_{ij}|C)$ for $C \in \{M, U\}$.

2.2. **Conditional independence.** If the comparison vector fields γ_{ij}^ℓ are independent across ℓ conditional on match status (i.e. if errors in name and address are not correlated conditional on match status then:

$$P(\gamma_{ij}|C) = \prod_{\ell=1}^{L} P(\gamma_{ij}^\ell|C)^{\gamma_{ij}^\ell}(1 - Pr(\gamma_{ij}^\ell|C))^{1-\gamma_{ij}^\ell} \qquad C \in \{M, U\} \tag{2}$$

if the $\{\gamma_{ij}^\ell\}_{\ell=1}^L$ are binary. This assumption reduces the number of parameters used to describe each mixture class from $2^L - 1$ (as there are $2^L - 1$ potential configurations of γ) to L

2.3. **Blocking.** If $n_1 \times n_2$ is large, the computations required to construct γ and iterate through different configurations of the data are unreasonable. It is therefore common to use blocking to limit the number of comparisons that are being studied.

2.4. **Prior information.** Let $p_{M\ell} = P(\gamma_{ij}^\ell|M)$, $p_{U\ell} = P(\gamma_{ij}^\ell|U)$. Assuming the conditional independence model (2) and global parameters that do not vary by block, a convenient prior distribution on the parameters is the product of independent Beta distributions,

$$\begin{aligned} p_M &\sim \text{Beta}(\alpha_M, \beta_M) \\ p_{M\ell} &\sim \text{Beta}(\alpha_{M\ell}, \beta_{M\ell}),\ \ell = 1, \ldots, L \\ p_{U\ell} &\sim \text{Beta}(\alpha_{U\ell}, \beta_{U\ell}),\ \ell = 1, \ldots, L \end{aligned}$$

Larsen (2012) notes that it is also possible to specify a prior distribution over the whole probability vector associated with the set of comparison vectors γ as two Dirichlet distributions:

$$Pr(\gamma|M) \sim \text{Dirichlet}(\delta_M)$$
$$Pr(\gamma|U) \sim \text{Dirichlet}(\delta_U)$$

I TRY BOTH in this paper?

2.5. **Non-binary comparisons/Levels.** Here I note that it is possible to model non-binary variables. (See Sadline), but that is beyond the current state of this project. Otherwise we can represent it in another way as in Sadinle (2017) but it's a pain. (Can include it)

2.6. **Gibbs sampler (Larsen, 2012).** For $i = 1, \dots, n_1 \in X_1$, $j = 1, \dots, n_2 \in X_2$ define

$$I_{ij} = \begin{cases} 1, & \text{if records } i \in X_1 \text{ and } j \in X_2 \text{ refer to the same entity;} \\ 0, & \text{otherwise} \end{cases}$$

Note that if we were to consider blocking, $I(\cdot)$ would be defined only for elements (i_s, j_s) belonging to the same block s.

If the match indicators I were known, the posterior distributions of individual parameters given values of the other parameters would be as follows:

$$p_M|I \sim \text{Beta}\left(\alpha_M + \sum_{(i,j)} I_{ij},\ \beta_M + \sum_{(i,j)} (1 - I_{ij})\right) \tag{3}$$

$$p_{M\ell}|I \sim \text{Beta}\left(\alpha_{M\ell} + \sum_{(i,j)} I_{ij}\gamma_{ij}^{\ell},\ \beta_{M\ell} + \sum_{(i,j)} I_{ij}(1 - \gamma_{ij}^{\ell})\right), \quad \ell = 1, \dots, L \tag{4}$$

$$p_{U\ell}|I \sim \text{Beta}\left(\alpha_{U\ell} + \sum_{(i,j)} (1 - I_{ij})\gamma_{ij}^{\ell},\ \beta_{U\ell} + \sum_{(i,j)} (1 - I_{ij})(1 - \gamma_{ij}^{\ell})\right), \quad \ell = 1, \dots, L \tag{5}$$

The posterior distribution of parameters can therefore be simulated via the following Gibbs Sampling scheme.

(1) Specify parameters for the prior distributions. Choose initial values of $\left(p_M^{(0)}, p_{M\ell}^{(0)}, p_{U\ell}^{(0)}\right)$ for $\ell = 1, \dots L$.

(2) Repeat the following steps numerous times until the distribution of draws has converged to the posterior distribution of interest:

   (a) Using the current values of $\left(p_M^{(k)}, p_{M\ell}^{(k)}, p_{U\ell}^{(k)}\right)$, draw $I_{ij}^{(k+1)}$ for each (i, j) candidate pair as an independent draw from a Bernoulli distribution with parameter

$$Pr\left(I_{ij}^{(k+1)} = 1 \middle| \gamma_{ij}\right) = Pr\left(M|\gamma_{ij}\right) = \frac{p_M^{(k)} Pr\left(\gamma_{ij}|M, p_{M\ell}^{(k)}\right)}{Pr\left(\gamma_{ij} \middle| p_M^{(k)}, p_{M\ell}^{(k)}, p_{U\ell}^{(k)}\right)} \tag{6}$$

where

$$Pr\left(\gamma_{ij}|M, p_{M\ell}^{(k)}\right) = \prod_{\ell=1}^{L} \left(p_{M\ell}^{(k)}\right)^{\gamma_{ij}^{\ell}} \left(1 - p_{M\ell}^{(k)}\right)^{1-\gamma_{ij}^{\ell}}$$

and the denominator is calculated according to (1) above.

(b) Draw a value of $p_M^{(k+1)}$ from (3).

(c) Draw values of $\{p_{M\ell}^{(k+1)}\}_{\ell=1}^{L}$ independently from (4).

(d) Draw values of $\{p_{U\ell}^{(k+1)}\}_{\ell=1}^{L}$ independently from (5).

(3) Stop once the algorithm has converged. Criteria for Convergence ARE X Y Z.

2.7. **Interpreting the results, one-to-one matching.** For a candidate pair (i, j), the posterior probability of a match is $P(I_{ij} = 1|p_M, p_{M\ell}, p_{U\ell}) = \frac{1}{K}\sum_k I_{ij}^{(k)}$. Options for designating matches are to (1) designate all candidate pairs exceeding a cutoff as matches, or (2) use a linear program to enforce one-to-one matching.

2.8. **Limitations.**

(1) There is no guarantee that the clusters will correspond to matches and non-matches. This is a more general criticism about using mixture models, which suffer from this identification problem. In practice, 3-component mixtures tend to work better, even though theoretically we would like two. Winkler (2002) mentioned conditions for the mixture model to give good results: the proportion of matches should be greater than 5%, the classes of matches and non-matches should be well-separated, typographical errors must be relatively low, there must be redundant fields that overcome errors in other fields, among others.

(2) Many-to-one matches can still happen unless the linear sum assignment is used. Even if mixture model is fitted with the one-to-one constraint, the FS decision rule alone may lead to many-to-many assignments. The linkage decision for the pair (i, j) not only depends on γ_{ij} but on the other pairs.

2.9. **Enforcing one-to-one assignment.** The FS decision rule does not enforce the maximum one-to-one assignment that is desirable in many economic applications. In practice, the optimal assignment of record pairs is obtained by solving the linear sum assignment problem:

$$\max_{\Delta} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} w_{ij}\Delta_{ij}$$

$$\text{subject to } \Delta_{ij} \in \{0, 1\}; \sum_{i=1}^{n_1} \Delta_{ij} \leq 1, \ j = 1, \ldots, n_2$$

$$\text{and } \sum_{j=1}^{n_2} \Delta_{ij} \leq 1, \ i = 1, \ldots, n_1$$

where the constraints ensure that Δ represents a bipartite matching. The output of this step is a bipartite matching that maximizes the sum of the weights w_{ij} among matched pairs, and the pairs that are not matched. Sadinle (2017) shows that this can be thought of as the MLE under the assumption that the comparison vectors are conditionally independent given the bipartite matching.

3. Testing the methods and results

I wrote a simple Python script that incoporates this naive Gibbs.

Things to do: Dirichlet, Betas, truncating the betas.

3.1. **Data.**

Description of DGP and simulations.

3.2. **More prior restrictions.**

It is expected that the probability of agreeing on an individual field of comparison is higher for matches than for nonmatches:

$$Pr\left(\gamma_{ij}^{\ell} = 1 |\ (i,j) \in M\right) \geq Pr\left(\gamma_{ij}^{\ell} = 1 |\ (i,j) \in U\right)$$

3.3. **Results.**

Some plots will go here.

3.4. **Disadvantages/Discussion.**

This method does not perform well, even when I impose strong priors and set initial parameter values equal to the truth. First, the posterior for p_M is heavily skewed toward 1 when I sample I using the formula in Step 1. This corresponds with high posterior probabilities of $I(a,b)$, which may imply large false positive matching rates if threshold is set too low.

This issue reflects the fact that updates to $(p_M, p_{M\ell}, p_{U\ell})$ depend on assignments of $I(a,b)$. The $sample_I$ function is too quick to assign matches. This may result from the fact that I use two clusters, and that once $p_{U\ell}$ probabilities get set low, the chain cannot recover. I test this issue by adding 1 to the denominator of the Bernoulli parameter in Step 1:

$$p \equiv Pr(\ I(a,b)^{(k+1)} = 1 \mid \gamma(a,b)) = Pr(\ M \mid \gamma(a,b)) = \frac{p_M^{(k)} Pr(\ \gamma(a,b) \mid M)}{Pr(\gamma(a,b)) + 1}$$

This change prevents p_M from converging to 1 ¡span style=”color:blue”¿(but I need to write more tests) ¡/span¿

My results are extremely sensitive to a choice of prior! Choosing the prior will be important to explore.

Could I sample from the joint distribution of $(p_M, p_{M\ell}, p_{U\ell})\ |I$? Could I model as Dirichlet?

Ultimately it is not worth the time and energy trying to fix this broken method so now I focus on the bipartite matching, which will fix many of these issues.

4. Bipartite Matchings

The main disadvantage of the mixture model approach is that it does not enforce one-to-one matching. Intuitively, if a one-to-one matching is desired, then the matching status of (i, j) should affect the matching status of (i, j'). Yet the mixture model approach does not allow for this.

This issue was noted by Larsen (2005) and Sadinle (2017). Formally, the parameter of interest is a bipartite matching, which can be represented compactly as a *matching labeling* $Z = (Z_1, Z_2, \dots, Z_{n_2})$ for the records in file X_2 such that

$$Z_j = \begin{cases} i, & \text{if records } i \in X_1 \text{ and } j \in X_2 \text{ refer to the same entity;} \\ n+1, & \text{if record } j \in X_2 \text{ does not have a match in } X_1 \end{cases}$$

4.1. Beta Prior for Bipartite Matchings Z.

This comes from Larsen (2005) and Sadinle (2017)/

Just as in the mixture model approach, the prior probability that $j \in X_2$ is:

$$I(Z_j \leq n_1) \overset{i.i.d}{\sim} \text{Bernoulli}(p_M)$$

where p_M represents the proporiton of matches expected a priori as a fraction of the smallest file X_2. Same as before, the hyperprior for p_M is:

$$p_M \sim \text{Beta}(\alpha_M, \beta_M)$$

The prior on p_M implies $n_{12}(Z) = \sum_{j=1}^{n_2} I(Z_j \leq n_1)$, the number of matches according to matching labeling Z is distributed as:

$$n_{12}(Z) \sim \text{Beta-Binomial}(n_2, \alpha_M, \beta_M)$$

after marginalizing over p_M.

Conditioning on $\{I(Z_j \leq n_1)\}_{j=1}^{n_2}$, all possible bipartite matchings are taken to be equally likely, so

$$Pr(Z \mid n_{12}) = \left(\frac{n_1!}{(n_1 - n_{12})!} \right)^{-1}$$

These conditions imply the joint prior over Z:

$$Pr(Z \mid \alpha_M, \beta_M) = \frac{(n_1 - n_{12}(Z))!}{n_1!} \frac{\text{Beta}(n_{12}(Z) + \alpha_M,\ n_2 - n_{12}(Z) + \beta_M)}{\text{Beta}(\alpha_M, \beta_M)}$$

4.2. **Gibbs sampler for bipartite matching (Larsen, 2005).**

(1) Pick an initial values of $p_M, p_{M\ell}, p_{U\ell}$ and a valid configuration of Z. Repeat the following until convergence:

   (a) Draw p_M from
   $$p_M \mid Z \sim \text{Beta}(\alpha_M + n_M(Z),\ \beta_M + n_2 - n_M(Z))$$
   Note this is same as before.

   (b) Draw $p_{M\ell}$ and $p_{U\ell}$ from their conditional distributions (same as before).

   (c) Use Metropolis-Hastings algorithm to draw values of Z and $n_M(Z)$ from their full conditional distributions.

(2) Stop when converged.

The only difference is step 3. We no longer draw matches according to $n_1 \times n_2$ independent Bernoulli random variables. The procedure used in this paper is exactly that from the appendix of Larsen (2005). It is quite long, so see Appendix for details.

5. Dicussion/Literature

Other option is to use training data to obtain the weights. Another is to apply mixture models to the comparison vectors directly. This latter method is favorable because in many contexts training data is not available, or creating a sufficiently large, representative training data set is costly.

Referee for Belin (93): "every gain which is achieved by a superior record linkage procedure must be justified by the cost of implementing that procedure".

The record linkage literature can be divided into two categories, broadly. The first are those who develop methods to perform matches that are in some way "optimal" – defined by reducing false match rates, or in terms of speed. The second, where this work will eventually fall, is studying how to incorporate uncertainty from the matching process into the subsequent analysis. Rarely are we interested in the match itself, but the analysis produced using matched data.

5.1. **Methods people.** Larsen (2005) shows how to use a hierarchical Bayesian model to allow for matching probability of agreeing on fields of information to vary by block (but share a hyperprior ($\log(\alpha_s/\alpha_s + \beta_s) \sim \mathcal{N}$). He also shows one-to-one restrictions. Now n_{12}, the number of matches is:
$$n_m \sim Binomial(n_2, p_m)$$
where $p_m \sim Beta(\alpha_m, \beta_m)$ as before. The prior distribution for the set of matches, is uniform over the space of possible matching configurations. Suggests doing linear sum assignment procedure at each step.

5.2. **Experimental.** Belin (1993): Performance of record-linkage procedure can depend on a number of factors, including:

(1) Choice of matching variables

(2) Choice of blocking variables

(3) Assignment of weights to agreement or disagreement on various matching variables

(4) Handling of close but not exact agreement between matching variables

(5) Handling of missing data in one or both of a pair of records

(6) Algorithm for assigning candidate matches

(7) Choice of cutoff weight above which record pairs will be declared matched

(8) The site or setting from which data are obtained

In the Bayesian, case we don't need to declare anyone matched, could calculate the probability of "model" (match) space explored via Geweke and then use this to calculate proper weights on reviewed matches?

5.3. **bias people.** Scheuren and Winkler (1993) and Lahiri and Larsen (2005) propose bias-corrected estimators of coefficients in a linear regression model given data from a probabilistically linked file.

Chipperfield et al. (2011) consider the analysis of linked binary variables.

Building on Chambers (2009), Kim and Chambers (2012a, 2012b) (referred to as KC hereafter) investigate the analysis of linked data using a more general set of models fitted using estimating equations.

Kim and Chambers (2012b) review recent development in inference for regression parameters using linked data.

Chipperfield and Chambers (2015) describes a bootstrap approach to inference using estimating equations that is based on probabilistically linked data where the linked data file is created under the 1-1 constraint. They say this of the previous papers:

> Linkage models form the key feature of all of the above approaches. The linkage model describes the probability that a record on one file is linked to each of the records on another file. For a linkage model to be useful, it must properly take into account how records were linked. SW and LL do not allow for 1-1 linkage, where every record on one file is linked to a distinct and different record on the other, or for linkage in multiple passes or stages, both of which are commonly used in probabilistic record linkage. In theory, KC allows for 1-1 linkage, but imposes strong constraints on the linkage model in order to do so. KC also requires a clerical sample to estimate the parameters of the linkage model, something which is not always available in practice and which itself can be subject to measurement errors.

In a Bayesian approach none of this necessary maybe.

6. Bipartite matching

Bayesian approaches of Fortini et al. (2001) and Larsen (2002, 2005, 2010) improve the mixture model implementation by properly treating the parameter of interest as a bipartite matching, which relaxes the assumption that record pairs' matching status is independent of one another. The setup is as follows:

Two sets: $X = \{x_1, \ldots, x_{n1}\}$ and $Y = \{y_1, \ldots, y_{n2}\}$. The goal is to find an assignment of items so that every item in X is matched to exactly one item in Y and no two items share the same match. An assignment corresponds to a permutation π where π is a one-to-one mapping (check?) $\{1, \ldots, n_1\} \to \{1, \ldots, n_2\}$ mapping each item in X to its match in Y. We define $\pi(i) = j$ to denote the index of a match $y_{\pi(i)} = y_j$ for an item x_i, and $\pi^{-1}(j) = i$ to denote the reverse (if it exists).

Uncertainty over assignments expressed as:

$$P(\pi|\theta) = \frac{1}{Z(\theta)} \exp(-E(\pi, \theta))$$

7. Priors

It is expected that the probability of agreeing on an individual field of comparison is higher for matches than for non-matches is $P(\gamma_k(a, b) = 1|(a, b) \in M) > P(\gamma_k(a, b) = 1|(a, b) \in U)$. And the probability of a match p_M is less than or equal to the minimum file size divided by the number of possible pairs.

8. Datasets

The techniques are tested first using a synthetic data generator developed by Christen and Pudjijono (2009) and Christen and Vatsalan (2013), and then with two real datasets from Enamorado (2018) and Enamorado, Fifield and Imai (2018).

(1) In the first empirical application I merge two datasets on local-level candidates for the 2012 and 2016 municipal elections in Brazil. Each dataset contains more than 450,000 observations with a perfectly-recorded unique identifier, the Brazilian individual taxpayer registration identification number (called the *Cadastro de Pessoas Físicas*). All other common identifiers are manually entered into the database, so they may contain errors.

(2) In the second application, I merge the 2016 American National Election Study(ANES) with a nationwide voter file containing over 160 million voter records.

9. Simulations

Description of the fake data.

9.1. **Sadinle (2017) Stuff.** What this section will include:

(1) Have plots for posterior of all M and U probabilities as produced by Sadinle procedure. Put on same graph?

(2) Posterior probability of matches? Try doing geweke?

(3) Plot of likelihood to show convergence

Files will look like

(1) For each data set, generates chain of bipartite matches Z, parameters for M and U probabilities, and writes these to files

(2) File that analyzes Z draws and gives posterior probability for pct correct, pct match, etc. across draws. Can also look at individual matches (will be useful for comparing the two methods)

(3) File that produces posterior distributions for m and u probabilities.

(4) Plot the likelihood to examine convergence properties.