BAYESIAN RECORD LINKAGE

RACHEL ANDERSON

1. Introduction

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

Ask any economist who works with survey or administrative data how to 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 by the automated merge processes in R and Stata. There is no one-size-fits-all solution to the multiple match problem, nor a formal theory about how these subjective matching decisions may introduce bias or uncertainty into estimation and inference in economic models.

This "merging" problem, defined formally as 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, merging 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 posterior probabilities over a match (are they consistent?). This is the problem that was introduced by Bo Honore (need a consistent probability of the match to ensure consistency of estimators). These weights can then be used to perform the analysis using standard econometric models.

The outline for this paper is as follows. First I will review some of the existing Bayesian techniques for record linkage, especially those from Larsen and Rubin (2001), Sadinle (2017), and Enamorado (2018). Then I will explore how to incorporate those weights in practice.

2. Literature

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.

2.1. **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.

2.2. **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?

2.3. **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.

Linkage model is a permutation matrix.

3. Setup (Sadinle, 2017)

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, so that the goal of record linkage is to identify which records in files X_1 and X_2 refer to the same entities.

Formally, the set of records coming from the two files can be represented as a *bipartite matching*. 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_{12}, so that $0 \leq n_{12} \leq n_2$.

One way to write the parameter of interest is as a *matching matrix* Δ of size $n_1 \times n_2$ whose (i, j)th entry is defined as:

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

A more compact representation used by Sadinle (2017) is a *matching labeling* $Z = (Z_1, Z_2, \ldots, Z_{n_2})$ for the records in file X_2 such that

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

I will use either representation interchangeably throughout the paper depending on convenience.

4. Probabilistic Record Linkage via Mixture Models

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):

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

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

4.1. **Mixture Models.** The probability of a particular configuration of the comparison vector γ_{ij} can be modeled as arising from a 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 among the matches and non-matches, respectively, and p_M and $p_U = 1 - p_M$ are the marginal probabilities of matches and unmatched pairs.

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

$$P(\gamma_{ij}|C) = \prod_{f=1}^{F} P(\gamma_{ij}^f|C)^{\gamma_{ij}^f}(1 - Pr(\gamma_{ij}^f|C))^{1-\gamma_{ij}^f} \qquad C \in \{M, U\}$$

if the $\{\gamma_{ij}^f\}_{f=1}^F$ are binary. Otherwise we can represent it in another way as in Sadinle (2017) but it's a pain.

4.2. **Prior Distributions.** The conditional independence assumption allows for a prior distribution on the parameters that is the product of independent Beta distributions,

$$p_M \sim Beta(\alpha_M, \beta_M)$$
$$P(\gamma_k(a,b) = 1 \mid M) \sim Beta(\alpha_{Mk}, \beta_{Mk}), \quad k = 1, \ldots, K$$
$$P(\gamma_k(a,b) = 1 \mid U) \sim Beta(\alpha_{Uk}, \beta_{Uk}), \quad k = 1, \ldots, K$$

Larsen (2005) notes that conditional independence across comparison vector fields may be relaxed by specifying a Dirichlet prior distribution on the whole probability vector associated with the set of comparison vectors γ_{ij}, however the dimension could be very large:

$$P(\gamma|C) \sim Dirichlet(\delta_C) \qquad C \in \{M, U\}$$

Alternatively, "training data" could inform the priors (Belin and Rubin).

4.3. **Estimation via Gibbs Sampling.** Mixture models typically use EM algorithm but here it's not necessary. Gibbs Sampler here.

4.4. **Restrictions on Parameters.**

4.5. **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.

4.6. **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.

5. 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:

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

6. 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.

7. 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.