BAYESIAN RECORD LINKAGE

RACHEL ANDERSON

1. Introduction

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.**

2.2. 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

4. Fellegi-Sunter Approach

4.1. Mixture Models.

Mixture models are useful for implementing the Fellegi-Sunter approach. Following Shalizi (ref here):

We say that the distribution f is a mixture of K component distributions $f_1, \dots, f_K$ if

$$f(x) = \sum_{k=1}^{K} \lambda_k f_k(x; \theta_k)$$

where λ_k are the mixing weights, such that $\lambda_k > 0$, $\sum_k \lambda_k = 1$. This means that the data can generated to the following procedure:

$$\begin{aligned} Z &\sim \text{Multinomial}(\lambda_1, \dots, \lambda_K) \\ X|Z &\sim f_Z \end{aligned}$$

where Z is a discrete random variable that says which component X is drawn from.

This is useful for record linkage using the FS approach because we assume there are two latent populations corresponding to matches (M) and non-matches (U), that are represented in the population of comparisons with proportions p_M and $p_U = 1 - p_M$ respectively.

Given $(i,j) \in M$ or $(i,j) \in U$, the comparison vector between two files i and j is

$$\gamma(i,j) \sim f_k, \quad k \in \{M, U\}$$

Assuming independent samples, the log likelihood for a generic mixture model for observations $(x_1, \ldots, x_n)$ is:

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{n} \log f(x_i, \theta) && (1) \\ &= \sum_{i=1}^{n} \log \sum_{k=1}^{K} \lambda_k f_k(x_i; \theta_k) && (2) \end{aligned}$$

The overall parameter vector of the model is thus $\theta = (\lambda_1, \ldots, \lambda_{K-1}, \theta_1, \theta_2, \ldots, \theta_K)$.

In record linkage, $x_i \overset{i.i.d.}{\sim} X$ and $y_j \overset{i.i.d.}{\sim} Y$, so $\gamma(x_i, y_j) \perp\!\!\!\perp \gamma(x_{i'}, y_j)$ for all $j, i' \neq i$, yet, $\gamma(x_i, y_j)$ may not be independent from $\gamma(x_i, y_{j'})$. This gives the likelihood:

$$\ell(\theta) \qquad (3)$$

4.2. **Mixture Model Estimation.** As shown in Shalizi (ref), maximizing the likelihood for a mixture model is like doing a weighted likelihood maximization, where the weight of each observation x_i depends on the cluster. This is seen by taking the derivative of (2) with respect to one parameter θ_j,

$$\begin{aligned} \frac{\partial \ell}{\partial \theta_j} &= \sum_{i=1}^{n} \frac{1}{\sum_{k=1}^{K} \lambda_k f_k(x_i; \theta_k)} \lambda_j \frac{\partial f_j(x_i; \theta_j)}{\partial \theta_j} \\ &= \sum_{i=1}^{n} \underbrace{\frac{\lambda_j f_j(x_i; \theta_j)}{\sum_{k=1}^{K} \lambda_k f_k(x_i; \theta_k)}}_{w_{ij}} \frac{\partial \log f_j(x_i; \theta_j)}{\partial \theta_j} \end{aligned}$$

The weight is the conditional probability that observation i belongs to cluster j :

$$w_{ij} = \frac{\lambda_j f_j(x_i; \theta_j)}{\sum_{k=1}^{K} \lambda_k f_k(x_i; \theta_k)} = \frac{P(Z = j, X = x_i)}{P(X = x_i)} = P(Z = j | X = x_i)$$

So if we try to estimate the mixture model, we're doing weighted maximum likelihood, with weights given by the posterior cluster probabilities (which depend on parameters $\lambda_1, \ldots, \lambda_K$ that we are trying to estimate). The EM algorithm (ref here for Rubin, etc.) makes estimation possible:

(1) Start with guesses about the mixture components $\theta_1, \theta_2, \ldots, \theta_K$ and the mixing weights $\lambda_1, \ldots, \lambda_K$.

(2) Until nothing changes very much:

    (a) Using the current parameter guesses, calculate the weights w_{ij} (E-step)

    (b) Using the current weights, maximize the weighted likelihood to get new parameter estimates (M-step)

(3) Return the final estimates for $\theta_1, \ldots, \theta_K, \lambda_1, \ldots, \lambda_K$ and clustering probabilities.

4.3. **Specifying mixture distributions.** The mixture distributions $f_k(x_i, \theta_k)$ need not be the same parametric family, nor do they even need be parametric. But in principal the different density mixtures could be any probabilistic model.

In the FS approach, agreement fields are conditionally independent. Makes sense to specify via Beta distributions. But could also specify a dirichlet distribution for each part...?

4.4. **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.

4.5. **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. 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.