Fastlink

Haocheng Qin, Kewei Xu, Zaolin Zhang, Chuhan Guo

Introduction

Methodology

Let's first describe the canonical probabilistic model of record linkage to demonstrate its properties. Materials are from the

original papaer: https://imai.fas.harvard.edu/research/files/linkage-app.pdf

and its online supplementary: https://imai.fas.harvard.edu/research/files/linkage-app.pdf

Let a latent mixing variable M_{ij} to indicate whether a pair of records (for the ith record in the data set A and the jth record in the data set B). Notate $\gamma(i,j)$ to be the distance between a pair, just as Fellegi and Sunter did[1], it's changebable but we just use Jaro-Winkler string distance here, which is commonly used[2]; and $\delta(i,j)$ to the missing indicator, to deal with missing data, then generally, we are estimating $\xi_{ij} := \Pr\left(M_{ij} = 1 \mid \delta(i,j), \gamma(i,j)\right)$. We can notice that, actually, without the missing indicator, Fastlink would be quite similar to Fellegi-Sunter. Even the assumptions are the same:

$$\begin{aligned}
&\gamma_k(i,j) \mid M_{ij} = m \overset{\text{indep.}}{\sim} \text{Discrete}\left(\pi_{km}\right); \\
&M_{ij} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\lambda)
\end{aligned}$$

But this can be actually relaxed, we will discuss it later. So addressing missing data would surelly be a significant property of Fastlink. Theoretically, it relies on a missing at random assumption (i.e. $\delta_k(i,j) \perp \gamma_k(i,j) \mid M_{ij}$), but it still works well under some other conditions. With some Bayes calculation (still similar to what Fellegi-Sunter did), actually we can obtain the exact probability to be:

$$\xi_{ij} = \frac{\lambda \prod_{k=1}^{K} \left(\prod_{\ell=0}^{L_k-1} \pi_{k\ell\ell}^{1\{\gamma_k(i,j)=\ell\}}\right)^{1-\delta_k(i,j)}}{\sum_{m=0}^{1} \lambda^m (1-\lambda)^{1-m} \prod_{k=1}^{K} \left(\prod_{\ell=0}^{L_k-1} \pi_{km\ell}^{1\{\gamma_\ell(i,j)=\ell\}}\right)^{1-\delta_k(i,j)}}$$

For the package, this is done by getPosterior(). This introduces us two main advantages:

Firstly, probabilistic models can quantify the uncertainty inherent in many merge procedures, offering a principled way to calibrate and account for false positives and false negatives. Also for post-merge analysis, such probability works as a good weight for merged variable, i.e. $X_i^* = \sum_{j=1}^{N_{\mathcal{B}}} \xi_{ij} X_j / \sum_{j=1}^{N_{\mathcal{B}}} \xi_{ij}$

Secondly, this proviade an easy way to compute. Intuitively, we can plug in the maximum likelihood estimation of λ and π here, which is

$$L_{com}(\lambda, \boldsymbol{\pi} \mid \boldsymbol{\gamma}, \boldsymbol{\delta}) \propto \prod_{i=1}^{N_A} \prod_{j=1}^{N_B} \prod_{m=0}^{1} \left\{ \lambda^m (1-\lambda)^{1-m} \prod_{k=1}^{K} \left(\prod_{\ell=0}^{L_k-1} \pi_{km\ell}^{\mathbf{1}\{\gamma_k(i,j)=\ell\}} \right)^{1-\delta_k(i,j)} \right\}^{\mathbf{1}\{M_{ij}=m\}}$$

which is hard to compute, but iteratively, we can apply EM (Expectation-Maximization) method[3] with

$$\lambda = \frac{1}{N_A N_B} \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \xi_{ij}$$

$$\pi_{km\ell} = \frac{\sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \mathbf{1}\{\gamma_k(i,j) = l)\Big\} \left(1 - \delta_k(i,j)\right) \xi_{ij}^m \left(1 - \xi_{ij}\right)^{1-m}}{\sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \left(1 - \delta_k(i,j)\right) \xi_{ij}^m \left(1 - \xi_{ij}\right)^{1-m}}$$

together with the calculated ξ_{ij} above. You can find the codes for EM calculations in “emlinkMARmov.R”, by codes you can find this package actually allows a prior for the hyperparameters, which is not shown above.

Actually we can find another function called “emlinklog.R” in the package, this is actually accommodating a more general pattern of interaction. This algorithm allows for the inclusion of interaction terms, that is what I said the assumptions can be relaxed, but without a prior, you can choose it by state “cond.indep = False” in the main function fastlink() according to your requirements, but we will not go through the details of this algorithm here.

Model Framework and Structures

Setup

The method involves merging two datasets, A and B, each containing NA and NB records respectively. They use K linkage variables for comparisons. The model defines an agreement vector $g(i,j)$ for each record pair (i,j), where $g_k(i,j)$ defines the similarity of the k-th variable between records i from A and j from B.

Model Formulation

- **Linkage Variables**: Uses Bernoulli random variables M_{ij} that identify whether a record pair (i,j) matches ($M_{ij} = 1$) or not ($M_{ij} = 0$). In other words, The model uses simple yes/no variables, represented mathematically as Bernoulli random variables M_{ij}. These variables help decide whether a pair of records (i,j) from two different datasets is a match ($M_{ij} = 1$) or not ($M_{ij} = 0$). Think of it as a sophisticated way of saying “these two records are talking about the same thing/person.”
- **Conditional Distributions**: Assumes conditional independence among linkage variables given the match status M_{ij}. In other words, each variable’s match status (like name, address) does not depend on each other after knowing whether the overall records match. This could enable the decomposition of the joint probability distribution into simpler, individual probabilities.
- **Handling Missing Data**: Utilizes a Missing At Random (MAR) framework to allow the omission of missing data in the probability calculations, which simplifies the likelihood function and enhances computational efficiency.

Algorithm and Computation

EM Algorithm

The parameter estimation is executed using the Expectation-Maximization (EM) algorithm. It starts with an initial guess, then repeatedly adjusts this guess aiming to improve the likelihood that the observed data came from the proposed model. This optimizes the observed-data likelihood function, which integrates over the probabilistic distributions of the linkage variables conditioned on the match hypotheses.

Blocking and Filtering

To reduce computational demands: - **Blocking**: To avoid comparing every record in one dataset with every record in another, which can be overwhelmingly time-consuming with large datasets, the model groups records into blocks based on shared characteristics (like all people with the same birth year), which greatly cuts down on unnecessary comparisons. - **Filtering**: Eliminates highly unlikely pairs from consideration early in the process, using thresholds based on calculated probabilities.

Scalability

The algorithm is designed to work efficiently even with very large datasets that contain millions of records. It uses parallel processing (splitting the work across multiple computer processors) and smart data structures to manage this, making it practical to run on a typical laptop without needing supercomputer resources.

Evaluation and Implementation

Simulation Studies

The model's robustness is tested through simulations that mimic real-world problems like incomplete data or errors in the data (measurement errors). These simulations help verify that the model can handle different types of common data issues effectively. The model is compared to traditional methods (like exact match), showing that it can handle complex, imperfect data more effectively and efficiently.

Package Realization

Repository: https://github.com/kosukeimai/fastLink

Example: https://imai.fas.harvard.edu/research/files/turnout.pdf

Implementation

Dataset: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/2NNA4L

Statistical Analysis Post-Merging

Uncertainty Quantification

The model quantifies the uncertainty in the merging process, allowing researchers to account for potential errors in subsequent analyses, which is critical for maintaining the integrity of research conclusions.

Post-Merge Analysis

Discusses methodologies for incorporating the probabilities of matches into regression analyses and other statistical procedures to adjust for the uncertainty inherent in the linkage process.

Contributions and Innovations

The model makes substantial contributions to the field of data management by providing: - A robust probabilistic framework that substantially outperforms traditional deterministic methods. Unlike older methods that just said 'yes' or 'no' to whether records match, this model calculates how likely it is that records match. This approach gives us a clearer picture and usually results in better performance. - Enhanced handling of missing data and the independence assumptions of linkage variables, which have been a significant limitation in earlier models. - Detailed documentation and an accessible implementation in R, which facilitates reproducible research and widespread adoption in the social sciences.

```r
suppressMessages(require("fastLink"))
suppressMessages(require("plyr"))
data <- read.delim("cces2016voterval.tab")
summary(data)
```

```
##       V101             merge_type    agreement_pattern    prob_match
##  Min.   :222168628   Min.   :1.000   Length:64600       Min.   :0.0000211
##  1st Qu.:302801850   1st Qu.:1.000   Class :character   1st Qu.:0.0173844
##  Median :303320104   Median :1.000   Mode  :character   Median :1.0000000
##  Mean   :303452665   Mean   :1.014                      Mean   :0.6659646
##  3rd Qu.:303923982   3rd Qu.:1.000                      3rd Qu.:1.0000000
##  Max.   :307210331   Max.   :2.000                      Max.   :1.0000000  
##  clerical_review     vote2016         vote2014         vote2012     
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :1.0000   Median :1.0000   Median :0.0000   Median :1.0000  
##  Mean   :0.5859   Mean   :0.7162   Mean   :0.4858   Mean   :0.5961  
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
##  vote2016_prob    vote2014_prob    vote2012_prob    vote2016_clerical
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   
##  Median :1.0000   Median :0.0000   Median :0.1620   Median :1.0000   
##  Mean   :0.5824   Mean   :0.4167   Mean   :0.4879   Mean   :0.5282   
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   
##  vote2014_clerical vote2012_clerical
##  Min.   :0.0000    Min.   :0.0000   
##  1st Qu.:0.0000    1st Qu.:0.0000   
##  Median :0.0000    Median :0.0000   
##  Mean   :0.3872    Mean   :0.4441   
##  3rd Qu.:1.0000    3rd Qu.:1.0000   
##  Max.   :1.0000    Max.   :1.0000
```

References

[1] I. P. Fellegi and A. B. Sunter, “A theory for record linkage,” *Journal of the American Statistical Association*, vol. 64, no. 328, pp. 1183–1210, 1969.

[2] M. A. Jaro, “Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida,” *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 414–420, 1989.

[3] W. E. Winkler, *Using the EM algorithm for weight computation in the fellegi-sunter model of record linkage.* US Bureau of the Census Washington, DC, 2000.