Population-Size Calibrated Bayes Estimate for Bipartite Record Linkage

[redacted]

Department of Statistical Science, Duke University, Durham, NC, USA

Introduction

- Bipartite record linkage wrestles with the problem of identifying the *same* individuals across two *different* databases, where no duplication are observed within each file.
- Many motivating applications require the linkage process to not only output point estimates for the final linkage structure, but also report the **induced population size estimate**.
- The current Bayes estimate of Sadinle (2017) induces **highly biased** population size estimates under various noisy scenarios and are not suited to applications where the population size is a parameter of interest.
- Our proposed methods: 1) Two-Stage Augmented Bayes (AB) and 2) F-Score Bayes are motivated by point estimates for the final linkage structure that are **well-calibrated** for population size under noisy scenarios.

Background

Coreference matrix

$C = [c_{i,j}]$ is the **co-reference matrix** of size $n_{\mathcal{A}} \times n_{\mathcal{B}}$, where $c_{i,j} = 1$ if records i and j are a match, and $c_{i,j} = 0$ otherwise.

Comparison Vector

$\gamma_{ij} = (\gamma_{ij}^1, \gamma_{ij}^2, \ldots, \gamma_{ij}^j)$ is the **comparison vector** between record i and record j, which encodes the degree of similarity between the share fields like name, age, sex, etc.

Bayes' Estimates

In practice, C is an unknown parameter estimated through record linkage. Adopting a Bayesian approach, the record linkage algorithm outputs a **posterior probability distribution** on the linkage structure C: $P(C|\gamma)$. To obtain a Bayes' Estimate, we seek to find $\hat{C}$ such that

$$\hat{C} = \arg\min_{\hat{C} \in \mathcal{C}} \mathbb{E}\left[L(C, \hat{C})\right] \tag{1}$$

Sadinle (2017)'s and AB's Loss Function:

$$L(C, \hat{C}) = \sum_i \sum_j L(C_{ij}, \hat{C}_{ij}) \tag{2}$$

F-Score Bayes Loss Function:

$$L_\beta^F(\hat{C}, C) = -\frac{(1+\beta^2)\sum_{i,j} \hat{c}_{i,j} c_{i,j}}{\beta^2 \sum_{i,j} c_{i,j} + \sum_{i,j} \hat{c}_{i,j}}. \tag{3}$$

Induced Bayes Estimate of Population Size

Any Bayes estimate of the final linkage structure $\hat{C}$ will induce a **population size estimate** defined as as,

$$\widehat{n_{12}} = \sum_j \sum_i \hat{C}_{ij} \tag{4}$$

Goal: we would like the true population size to equal (or close to) to the induced estimated population size i.e. $\widehat{n_{12}} \approx n_{12}$.

Characterization of Noise

1. Homonomy Rate for Field f: For field f with l levels, the Homonomy Rate is the proportion of non-link comparisons that agree on field f but are non-links, where d_{ijf} is the disagreement indicator:

$$H_f = \frac{\sum_{j=1}^{n_{\mathcal{B}}} \sum_{i=1}^{n_{\mathcal{A}}} I(C_{ij} = 0) I(d_{ijf} = 0)}{\sum_{j=1}^{n_{\mathcal{B}}} \sum_{i=1}^{n_{\mathcal{A}}} I(C_{ij} = 0)}$$

2. Variation Rate for Field f: For field f with l levels, the Variation Rate is the proportion of link comparisons that disagree on field f but are true links:

$$V_f = \frac{\sum_{j=1}^{n_{\mathcal{B}}} \sum_{i=1}^{n_{\mathcal{A}}} I(C_{ij} = 1) I(d_{ijf} = 1)}{\sum_{j=1}^{n_{\mathcal{B}}} \sum_{i=1}^{n_{\mathcal{A}}} I(C_{ij} = 1)}$$

Table 1: Different types of Noise in Name

Name	True ID	Homonomy	Variation
John Smith	1	1	1
John Smith	2	1	0
John J. Smith	1	1	1
John J. Smith	3	1	0
Mike Amiri	4	0	0

Shortcoming of Sadinle (2017)

Corollary 2: Sufficient Condition for a Non-Match:
Under the unity cost assumption applied in (2), A sufficient condition for a non-match for record j is if,

$$1 - \sum_i P(C_{ij} = 1|\gamma) \geq \max_{i:i \leq n_{\mathcal{A}}} P(C_{ij} = 1)$$

Table 2: Moderate Noise for Linkage Decision for individual j

i	$P(C_{ij} = 1\|\gamma_{ij})$
1	0.25
2	0.25
3	0.10
4	0.09
5	0.01
non-match	0.30

Table 3: Moderate Noise for Linkage Decision for individual j

i	$P(C_{ij} = 1\|\gamma_{ij})$
1	0.05
2	0.05
3	0.05
...	0.05
$n_{\mathcal{A}}$	0.05
non-match	0.10

Method 1: Two-Stage Augmented Bayes (AB)

Setup

- $A_{Z_j} = \{P(C_{ij} = 1|\gamma_{ij}) | i \leq n_1\}$ be the set of all posterior probability of a match for C_{ij}
- k is the number of augmented units for comparison
- T is the lower bound on the probability for declared links

Modified Sufficient Condition for a Non-Match:
We will declare a non-match for record j in dataset 2 ($\hat{C}_{ij} = 0, \forall i$) if

$$P(C_{ij} = 0|\gamma^{\mathbf{obs}}) \geq \max_{A'_{Z_j} \subseteq A_{Z_j}: |A'_{Z_j}| = k} \sum_{a \in A'_{Z_j}} a \tag{5}$$

Algorithm

Stage 0: Pick appropriate k and T.
Stage 1: Determine non-links from (5).
Stage 2: Run Linear Sum Assignment Problem algorithm (LSAP) on remaining links, conditioned on the links having a match in dataset 1.
Post-Processing: For any chosen link i^* such that $P(C_{ij} = 1) \leq T$, and declare j to be a non-link.

Method 2: F-Score Bayes

Setup

Proposition 3: Calibrated Population Size:
Let $P = \widehat{n_{12}}$ be the total number of predicted links, $T = n_{12}$ be the total number of true links, and TP is the true positive links. If Precision = Recall, then $T = P$.
Weighted F Score:

$$F_\beta(\hat{C}, C) = \frac{(1+\beta^2)\sum_{i,j} \hat{c}_{i,j} c_{i,j}}{\beta^2 \sum_{i,j} c_{i,j} + \sum_{i,j} \hat{c}_{i,j}}. \tag{6}$$

Bayes Estimator:

$$\hat{C}_{\text{Bayes}} = \arg\min_{\hat{C} \in \mathcal{C}} \mathbb{E}\left[L_\beta^F(\hat{C}, C)\right], \tag{7}$$

Algorithm

$$\hat{C}_{\text{Bayes}} = \arg\max_{k \in \mathbb{N}} \arg\max_{\hat{C} \in \mathcal{C}, \sum_{i,j} \hat{c}_{i,j} = k} \mathbb{E}\left[F_\beta(\hat{C}, C)\right] \tag{8}$$

$$= \arg\max_{k \in \mathbb{N}} \arg\max_{\hat{C} \in \mathcal{C}, \sum_{i,j} \hat{c}_{i,j} = k} \sum_{i,j} \hat{c}_{i,j} \mathbb{E}\left[\frac{(1+\beta^2) c_{i,j}}{\beta^2 \sum_{i,j} c_{i,j} + k}\right]. \tag{9}$$

For a given $k \in \mathbb{N}$, the inner optimization problem in (9) can be solved as a linear sum assignment(LSAP) problem with the constraint of k links using a simple modification of weight matrix in LSAP (see appendix A.2 in report).

Simulation Results

Table 4: Table of Performance for Moderate Noise Scenario

	Sadinle (2017)	2-Stage Augmented Bayes	F-Score Bayes
Misclassification #	21	16	**16**
2.5%	35	35	35
97.5%	50	50	50
True Population Size	36	36	36
Induced Population Size #	24	41	**40**

Table 5: Table of Performance for High Noise Scenario

	Sadinle (2017)	2-Stage Augmented Bayes	F-Score Bayes
Misclassification #	36	36	**26**
2.5%	1	1	1
97.5%	49	49	49
True Population Size	36	36	36
Induced Population Size #	0	0	**39**

Acknowledgement

This poster summarizes a forthcoming manuscript that will be produced jointly by Eric Bai and Statistical Science Ph.D. candidate Olivier Binette.

References

[1] Sadinle, M. Bayesian estimation of bipartite matchings for record linkage, *Journal of the American Statistical Association*

MSS Portfolio Completion Exercise