Modern Bayesian Record Linkage: Some Recent Developments and Open Challenges

Presented by Rebecca Steorts on Thursday 7th July 2016

Dayi Fang

Department of Statistical Science
Duke University

September 27 2016

Outline

1. Introduction to Record Linkage
    - Motivation and Examples
    - History and Recent Developments
    - Two major methods and problems

2. Bayesian Methods
    - Advantages and disadvantages
    - Two interesting papers

Outline

1. Introduction to Record Linkage
   - Motivation and Examples
   - History and Recent Developments
   - Two major methods and problems

2. Bayesian Methods
   - Advantages and disadvantages
   - Two interesting papers

Motivation and Examples

- Basic idea: remove duplicated administrative, medical, or other type of records.
- In practise: link information that under different tags

Outline

1. Introduction to Record Linkage
    - Motivation and Examples
    - History and Recent Developments
    - Two major methods and problems
2. Bayesian Methods
    - Advantages and disadvantages
    - Two interesting papers

History and Recent Developments

- Genetics: A Theory for Record Linkage (1969).
- Health, government, privacy, Bayesian methods (1980 - 2000).
- Modern Bayesian methods, machine learning, and clustering (2000 - 2016).

Outline

1. Introduction to Record Linkage
   - Motivation and Examples
   - History and Recent Developments
   - Two major methods and problems

2. Bayesian Methods
   - Advantages and disadvantages
   - Two interesting papers

Two major methods and problems

- Hand matching (scalability, cost).
- Fellegi and Sunter (contains transitive closures, but is still not scalable).
- Two major concerns: scalability and the level of interest.

Outline

1. Introduction to Record Linkage
   - Motivation and Examples
   - History and Recent Developments
   - Two major methods and problems

2. Bayesian Methods
   - Advantages and disadvantages
   - Two interesting papers

Advantages and disadvantages

- Bayesian methods can provide exact uncertainty from the linkage process.
- Bayesian methods are hard to generalize for multiple record linkage.
- Most methods do not scale well.

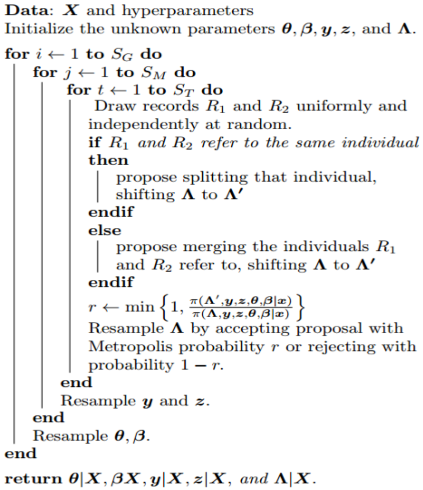
Outline

1. Introduction to Record Linkage
   - Motivation and Examples
   - History and Recent Developments
   - Two major methods and problems

2. Bayesian Methods
   - Advantages and disadvantages
   - Two interesting papers

A Bayesian Approach to Graphical Record Linkage and De-duplication

- Split and MErge REcord linkage and De-duplication (SMERED)



```
Data: $\boldsymbol{X}$ and hyperparameters
Initialize the unknown parameters $\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{z}$, and $\boldsymbol{\Lambda}$.
for $i \leftarrow 1$ to $S_G$ do
    for $j \leftarrow 1$ to $S_M$ do
        for $t \leftarrow 1$ to $S_T$ do
            Draw records $R_1$ and $R_2$ uniformly and
            independently at random.
            if $R_1$ and $R_2$ refer to the same individual
            then
                propose splitting that individual,
                shifting $\boldsymbol{\Lambda}$ to $\boldsymbol{\Lambda'}$
            endif
            else
                propose merging the individuals $R_1$
                and $R_2$ refer to, shifting $\boldsymbol{\Lambda}$ to $\boldsymbol{\Lambda'}$
            endif
            $r \leftarrow \min\left\{1, \frac{\pi(\boldsymbol{\Lambda'}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta} | \boldsymbol{x})}{\pi(\boldsymbol{\Lambda}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta} | \boldsymbol{x})}\right\}$
            Resample $\boldsymbol{\Lambda}$ by accepting proposal with
            Metropolis probability $r$ or rejecting with
            probability $1 - r$.
        end
        Resample $\boldsymbol{y}$ and $\boldsymbol{z}$.
    end
    Resample $\boldsymbol{\theta}, \boldsymbol{\beta}$.
end
return $\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{\beta}\boldsymbol{X}, \boldsymbol{y}|\boldsymbol{X}, \boldsymbol{z}|\boldsymbol{X}$, and $\boldsymbol{\Lambda}|\boldsymbol{X}$.
```

A Comparison of Blocking Methods for Record Linkage

- Simple Alternatives to Blocking: with the knowledge of the types of errors that are unlikely happened for a certain field or a combination of them, we can identify a pair of records as a non-match when it has strong disagreements in a combination of fields.
- Cluster-Based Blocking: based on the idea that the records in a cluster should be similar, selecting good candidate pairs for linkage
- Locality-Sensitive Hashing (LSH) based methods: LSH uses all of the information contained in each record to build manageably small blocks. High speed and high recall rate, low precision and a lot of flase positive.