

Causal Inference & Record Linkage

Kayané Robach



EPIDEMIOLOGY AND
DATA SCIENCE



1 Overview

2 Toy example

3 Record linkage recipe

4 A flexible model for record linkage

5 Simulations

6 Further work

Motivations

- How to merge multiple data sources? Record Linkage

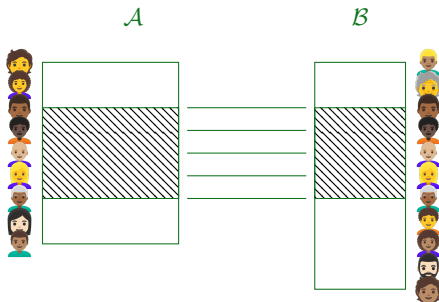


Figure: Illustration of the record linkage task.

Motivations

- What about causality?

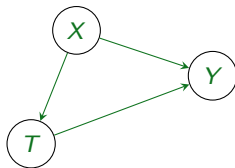


Figure: Illustration of a simple causal model.

Motivations

- you are interested in some causal effect
- covariates and treatment are in one data set from a first study \mathcal{A}
- outcomes are in another one from a second study \mathcal{B}
for e.g. long-term outcomes studies, survival data

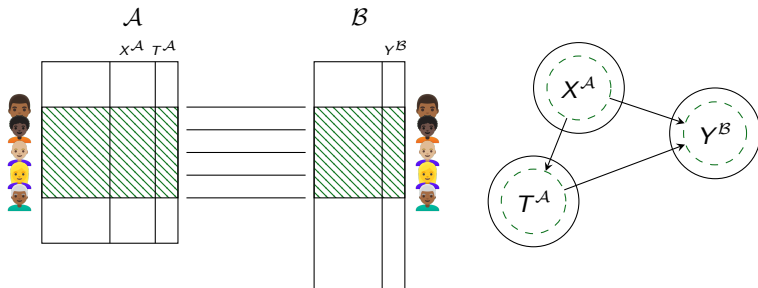


Figure: Illustration of a causal model study on matched pairs from a record linkage procedure.

Case study

Netherlands Perinatal Registry by Perined

- Study the risk of preterm birth using data from past deliveries
- Netherlands Perinatal Registry (PRN) collected by Perined with approximately 500,000 observations per data source

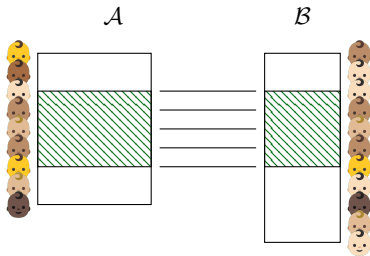


Figure: Illustration of the PRN problem where data are collected at the scope of deliveries without unique identifier for the mother.

- Registrations are at the delivery level and not the mother level
- First challenge is about matching babies from the same mother

1 Overview

2 Toy example

3 Record linkage recipe

4 A flexible model for record linkage

5 Simulations

6 Further work

Red thread example

Replenish your family tree

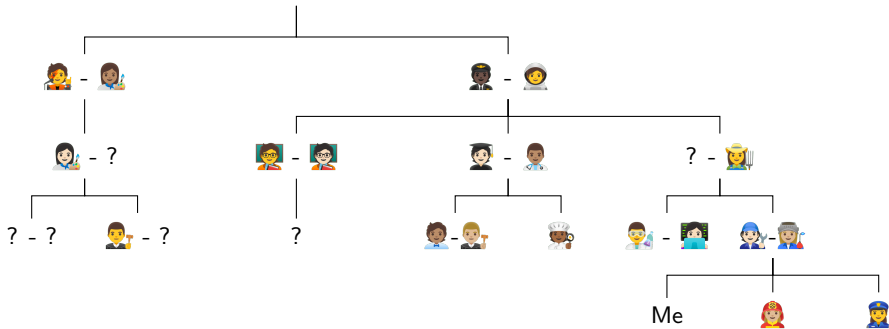



Figure: Toy illustration for the family tree task.

Red thread example

Replenish your family tree

- Start with:
 - list of names,
 - some basic information about your relatives:
 - the region they lived in,
 - their occupation,
 - ...

 : many people with the same names as your ancestors who may also have changed their names due to marriage, or other reasons

→ Record linkage uses probabilistic knowledge to match statistical information to reconstruct the portrait of your elders

Red thread example

Replenish your family tree

To replenish your family tree and understand the death circumstances of your forebears you need:

- 'partially' identifying variables (common to the multiple sources):
 - names,
 - birth date,
 - country,
 - sex assigned at birth,
 - ...
- some variables of interest:
 - X_j : the marital status, the number of children, the medical background
 - Y_j : the death circumstances

Red thread example

Replenish your family tree

- \mathcal{A} and \mathcal{B} , two datasets
- $\mathcal{A} = \{(\mathbf{G}_i^{\mathcal{A}}, X_i); i \in \llbracket 1, n_{\mathcal{A}} \rrbracket\}$ with $n_{\mathcal{A}}$ records
- $\mathcal{B} = \{(\mathbf{G}_j^{\mathcal{B}}, Y_j); j \in \llbracket 1, n_{\mathcal{B}} \rrbracket\}$ with $n_{\mathcal{B}}$ records

\mathcal{A}

Name	Family name	birth year	sex	children	blood
Kayané	Robachow	1899	F	0	0-
Stéphanie	van der Pas	1897	F	5	A+
Michel	Hof	1892	F	1	B-
Stephanie	Pas	1891	M	0	A+
Mark	Wiel	1891	M	2	AB-

\mathcal{B}

Family name	birth year	sex	death
Robach	1899	F	cancer
Pas	1891	F	old age
Hof	1892	M	murdered

Figure: Toy example of uncertain linkage scenario where the partially identifying variables are the family name, the birth year and the sex.

1 Overview

2 Toy example

3 Record linkage recipe

4 A flexible model for record linkage

5 Simulations

6 Further work

Work in progress

- Recent work develop
 - 2 stages models [Wortman and Reiter, 2018],
 - bayesian joint model [Tancredi and Liseo, 2011, Sadinle, 2017, Guha et al., 2022], for inference and record linkage
- We are working on a simple setting for record linkage (paper ongoing)
- Our model is flexible
 - can incorporate information useful for causal inference,
 - relies on tangible assumptions,
 - provides maximum likelihood estimates

A short manual of record linkage

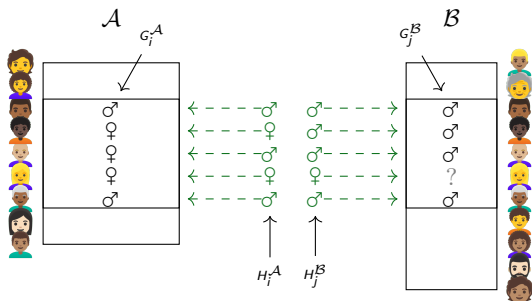


Figure: Illustration for the record linkage task.

- Model the **data generating process**
 - no reduction of the information
 - few possible linkages given the true registered values H_i^A, H_j^B
- Take linkage decisions independent of each other
- Incorporate the one-to-one assignment constraint

A short manual of record linkage

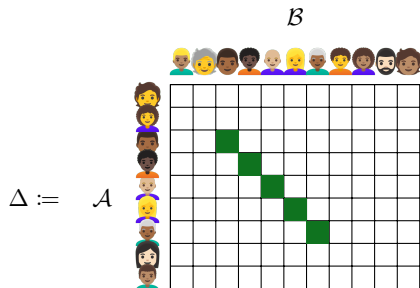


Figure: Illustration for the true linkage matrix Δ we would like to estimate.

- 1 Overview
- 2 Toy example
- 3 Record linkage recipe
- 4 A flexible model for record linkage**
- 5 Simulations
- 6 Further work

New model

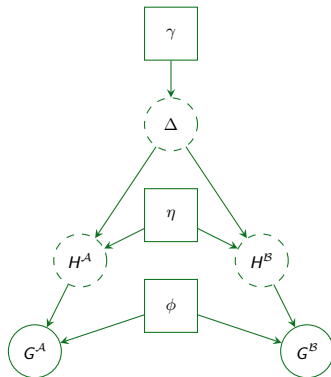


Figure: Probabilistic graphical model for the decomposition of the data generating process illustrating the record linkage problem. Circles indicates random variables while squares are reserved for parameters. Dotted lines indicates unobserved latent variables and solid lines observables.

Registered values

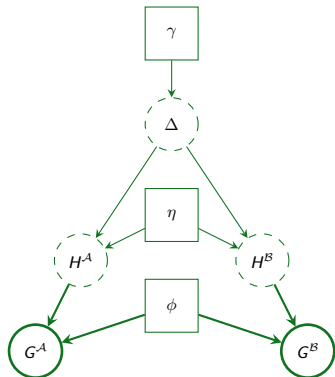


Figure: Probabilistic graphical model for the record linkage problem.

The registered partially identifying variables: G^A , G^B and their model^a

$$\mathbb{P}\left(G_i^A = a \mid H_i^A = b; \phi\right) =$$

$$\mathbb{P}\left(G_j^B = a \mid H_j^B = b; \phi\right)$$

where we can incorporate

- missing registered values,
- mistakes in registered values (compared to the truth),
- typos in registration (slightly different from the truth),

^aObservations in \mathcal{A} and in \mathcal{B} comes from the same super population, hence the same distribution for both set of individuals.

Latent true values

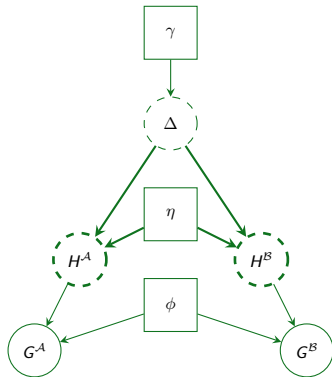


Figure: Probabilistic graphical model for the record linkage problem.

The latent partially identifying variables H^A , H^B and

- a model for linked records
 $\mathbb{P}(H_i^A = a, H_j^B = b \mid \Delta_{i,j} = 1; \eta),$
- a model for non-linked records^a
 $\mathbb{P}(H_i^A = a \mid \sum_{j=1}^{n_B} \Delta_{i,j} = 0; \eta) =$
 $\mathbb{P}(H_j^B = a \mid \sum_{i=1}^{n_A} \Delta_{i,j} = 0; \eta)$

^aObservations in \mathcal{A} and in \mathcal{B} comes from the same super population, hence the same distribution for both set of individuals.

Matching matrix

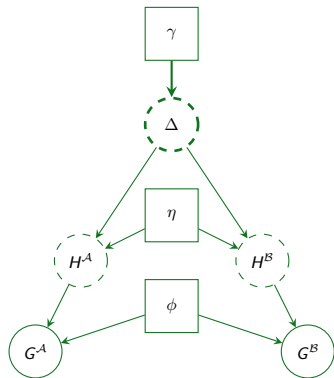


Figure: Probabilistic graphical model for the record linkage problem.

Our parameter of interest:

$$\Delta = \begin{pmatrix} \Delta_{1,1} & \Delta_{1,2} & \cdots & \Delta_{1,n_B} \\ \Delta_{2,1} & \Delta_{2,2} & \cdots & \Delta_{2,n_B} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta_{n_A,1} & \Delta_{n_A,2} & \cdots & \Delta_{n_A,n_B} \end{pmatrix}$$

with its definition set:

$$D = \left\{ \Delta : \Delta_{i,j} \in \{0, 1\}, \right.$$

$$\left. \sum_{i=1}^{n_A} \Delta_{i,j} \leq 1 \text{ for all } j \in \llbracket 1, n_B \rrbracket \right.$$

$$\left. \text{and } \sum_{j=1}^{n_B} \Delta_{i,j} \leq 1 \text{ for all } i \in \llbracket 1, n_A \rrbracket \right\}$$

and its model: $\mathbb{P}(\Delta; \gamma)$

- 1 Overview
- 2 Toy example
- 3 Record linkage recipe
- 4 A flexible model for record linkage
- 5 Simulations**
- 6 Further work

Simulations

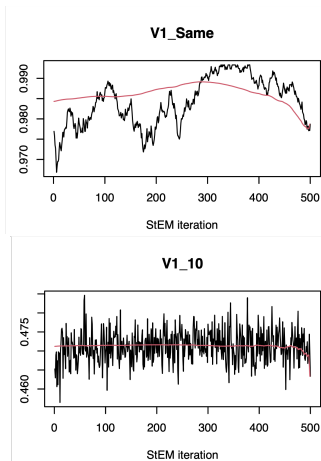


Figure: Convergence of one parameter for the partially identifying variables (probability of agreement) on the top. Convergence of one parameter for one categorical value of the true latent variables (probability of having a value equal to 10) on the bottom.

Simulations

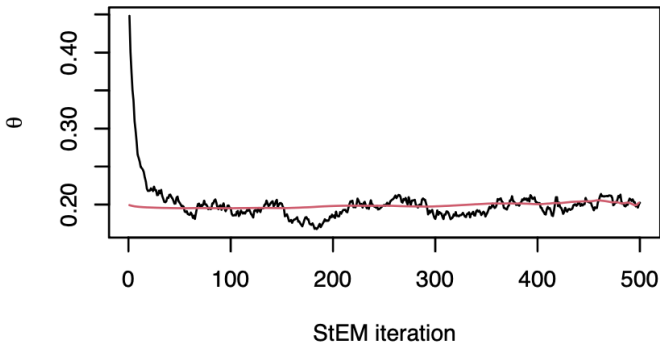


Figure: Convergence of the probability to have a link (truth: 20% of the smallest file) in a stochastic EM algorithm with 500 iterations and 100 iterations of the Gibbs sampler to draw the true values and the corresponding possible linkage matrices.

Simulations

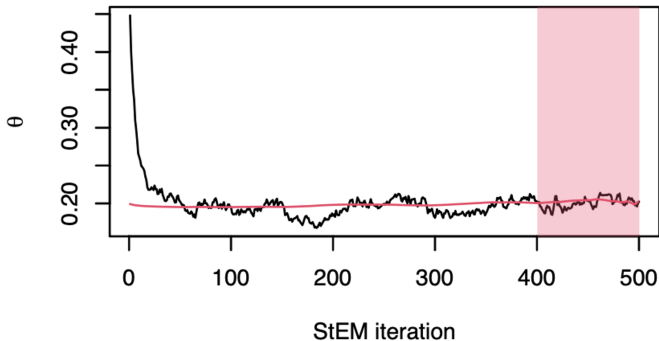


Figure: Convergence of the probability to have a link (truth: 20% of the smallest file) in a stochastic EM algorithm with 500 iterations and 100 iterations of the Gibbs sampler. The last 100 iterations are used to build our estimate of Δ .

Simulations

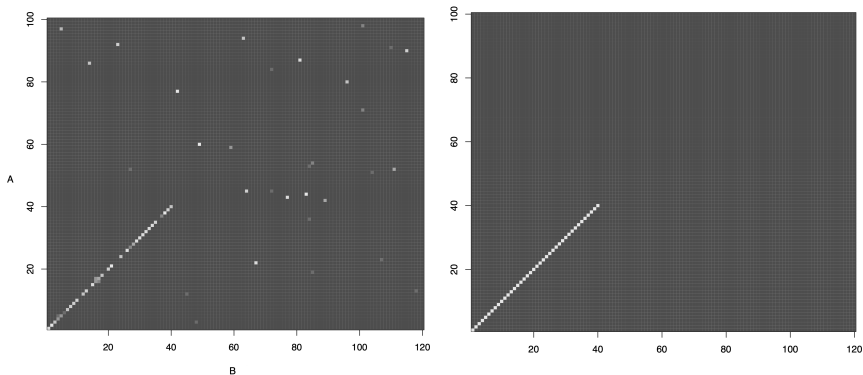


Figure: Result of the estimation of Δ compared to the truth for a stochastic EM with 500 iterations and 100 iterations of the Gibbs sampler. High probability of a match are in white.

1 Overview

2 Toy example

3 Record linkage recipe

4 A flexible model for record linkage

5 Simulations

6 Further work

Further work

- The flexibility of the model allows us to add knowledge in $\mathbb{P}(\Delta)$ useful for inference
- The likelihood estimate
 - enables to compute likelihood ratios and compare efficiency with other methods,
 - is unbiased
- Unlike recent Bayesian approach, our model is scalable to big datasets



Guha, S., Reiter, J. P., and Mercatanti, A. (2022).
Bayesian causal inference with bipartite record linkage.
Bayesian Analysis, 17(4):1275 – 1299.



Sadinle, M. (2017).
Bayesian estimation of bipartite matchings for record linkage.
Journal of the American Statistical Association, 112(518):600–612.



Tancredi, A. and Liseo, B. (2011).
A hierarchical bayesian approach to record linkage and population size problems.
The Annals of Applied Statistics, 5(2B).



Wortman, J. H. and Reiter, J. P. (2018).
Simultaneous record linkage and causal inference with propensity score
subclassification.
Statistics in Medicine, 37(24):3533–3546.

If you have multiple datasets you would like to merge to do inference, call us!

Thank You!