A STochastic Expectation Maximisation approach to Record Linkage

Kayané Robach, S. L. van der Pas, M. A. van de Wiel and M. H. Hof

Amsterdam UMC
Universitair Medische Centra

EPIDEMIOLOGY AND DATA SCIENCE

EDS

Bigstatistics

December 12, 2024

International Biometric Conference

IBC

Contents

Stochastic EM

The Gaussian mixture problem

$$y_1, \ldots, y_n \text{ i.i.d. obs. } p_{\boldsymbol{\theta}}(\boldsymbol{y}) = \sum_{k=1}^{\kappa} \omega_k \cdot \phi(\boldsymbol{y}; \mu_k, \Sigma_k),\ \boldsymbol{\theta}_k = \{\omega_k, \mu_k, \Sigma_k\}$$

The Gaussian mixture problem

$y_1, \ldots, y_n$ i.i.d. obs. $p_{\boldsymbol{\theta}}(\boldsymbol{y}) = \sum_{k=1}^{\kappa} \omega_k \cdot \phi(\boldsymbol{y}; \mu_k, \Sigma_k)$, $\boldsymbol{\theta}_k = \{\omega_k, \mu_k, \Sigma_k\}$

MLE $\hat{\boldsymbol{\theta}}_{ML}$ maximises
$$\begin{aligned}
\sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(y_i) &= \sum_{i=1}^{n} \log \sum_{k=1}^{\kappa} \omega_k \cdot \phi(y_i; \mu_k, \Sigma_k) \\
&= \sum_{i=1}^{n} \log \sum_{z_i} p_{\boldsymbol{\theta}}(y_i, z_i) \\
&= \sum_{i=1}^{n} \log \mathbb{E}_{p_{\boldsymbol{\theta}^t}(\cdot|y_i)} \left[\frac{p_{\boldsymbol{\theta}}(y_i, z_i)}{p_{\boldsymbol{\theta}^t}(z_i|y_i)} \right] \\
&\geq \sum_{i=1}^{n} \mathbb{E}_{p_{\boldsymbol{\theta}^t}(\cdot|y_i)} \left[\log \frac{p_{\boldsymbol{\theta}}(y_i, z_i)}{p_{\boldsymbol{\theta}^t}(z_i|y_i)} \right]
\end{aligned}$$

The Gaussian mixture problem

$y_1, \ldots, y_n$ i.i.d. obs. $p_{\boldsymbol{\theta}}(\boldsymbol{y}) = \sum_{k=1}^{\kappa} \omega_k \cdot \phi(\boldsymbol{y}; \mu_k, \Sigma_k)$, $\boldsymbol{\theta}_k = \{\omega_k, \mu_k, \Sigma_k\}$

MLE $\hat{\boldsymbol{\theta}}_{ML}$ maximises
$$
\begin{aligned}
\sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(y_i) &= \sum_{i=1}^{n} \log \sum_{k=1}^{\kappa} \omega_k \cdot \phi(y_i; \mu_k, \Sigma_k) \\
&= \sum_{i=1}^{n} \log \sum_{z_i} p_{\boldsymbol{\theta}}(y_i, z_i) \\
&= \sum_{i=1}^{n} \log \mathbb{E}_{p_{\boldsymbol{\theta}^t}(\cdot|y_i)} \left[\frac{p_{\boldsymbol{\theta}}(y_i, z_i)}{p_{\boldsymbol{\theta}^t}(z_i|y_i)} \right] \\
&\geq \sum_{i=1}^{n} \mathbb{E}_{p_{\boldsymbol{\theta}^t}(\cdot|y_i)} \left[\log \frac{p_{\boldsymbol{\theta}}(y_i, z_i)}{p_{\boldsymbol{\theta}^t}(z_i|y_i)} \right]
\end{aligned}
$$

EM maximises this observed data log-likelihood lower bound

StEM algorithm

The Expectation Maximisation method is introduced to iteratively compute maximum likelihood estimates from incomplete data, (Dempster et al., 1977; Wu, 1983; Delyon et al., 1999)

StEM algorithm

The Expectation Maximisation method is introduced to iteratively compute maximum likelihood estimates from incomplete data, (Dempster et al., 1977; Wu, 1983; Delyon et al., 1999)

Sometimes, the **M-step** is not explicit $\rightarrow$ all shots are allowed (as long as convergence is ensured)

StEM algorithm

The Expectation Maximisation method is introduced to iteratively compute maximum likelihood estimates from incomplete data, (Dempster et al., 1977; Wu, 1983; Delyon et al., 1999)

Sometimes, the **M-step** is not explicit $\rightarrow$ all shots are allowed (as long as convergence is ensured)

When the **E-step** is too difficult to derive, one needs to approximate the bound $\rightarrow$ we can sample latent data from $p_{\boldsymbol{\theta}^t}(\cdot|y)$

Stochastic **EM**, (Celeux and Diebolt, 1986)

StEM algorithm

The Expectation Maximisation method is introduced to iteratively compute maximum likelihood estimates from incomplete data, (Dempster et al., 1977; Wu, 1983; Delyon et al., 1999)

Sometimes, the **M-step** is not explicit $\rightarrow$ all shots are allowed (as long as convergence is ensured)

When the **E-step** is too difficult to derive, one needs to approximate the bound $\rightarrow$ we can sample latent data from $p_{\boldsymbol{\theta}^t}(\cdot|y)$

Stochastic **EM**, (Celeux and Diebolt, 1986)

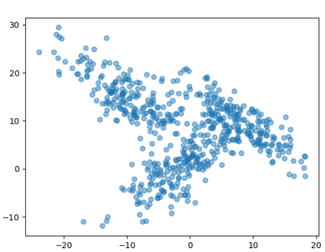
For a mixture model, this variant allows to identify the unknown number of clusters, and avoid convergence towards local maxima

Stochastic approach to EM

$z_1, \ldots, z_n \in \{1, \ldots, \kappa\}$ i.i.d. latent, $y_i | z_i = k, \boldsymbol{\theta}_k \sim \mathcal{N}(\mu_k, \Sigma_k)$

Expectation compute the cluster assignments

Maximisation adjust the cluster properties $\boldsymbol{\theta}_k = \{\omega_k, \mu_k, \Sigma_k\}$



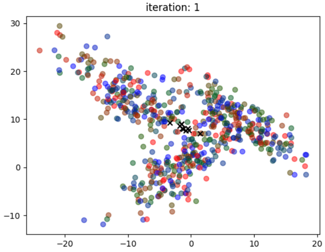


Figure 1: The StEM fitting a Gaussian Mixture

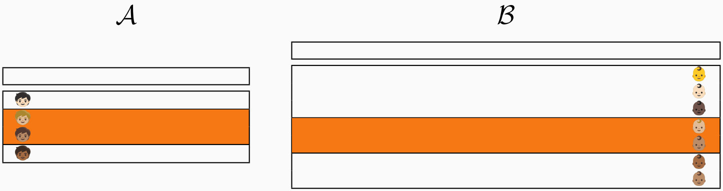
Record Linkage task

A motivational example

The Netherlands Perinatal Registry gathers about 96% of all deliveries

We could study the risk of pre-term birth using characteristics of the mother and data from past deliveries

Data are at the scope of the babies, family portraits need to be assembled



A motivational example

Make use of 'partially identifying variables' *postal code, birth date*

Cluster records according to *non-linked from $\mathcal{A}$, non-linked from $\mathcal{B}$, linked common to $\mathcal{A}$ and $\mathcal{B}$*

The true linkage structure is latent

$\mathcal{A}$

zipcode	delivery date	pre-term
1012GL	28-06-2021	yes
1112XJ	13-04-2019	no
8043VD	14-10-2015	yes
3572TC	03-08-2008	yes

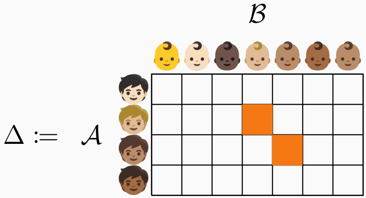
$\mathcal{B}$

Age	ART	zipcode	delivery date	pre-term	past delivery
25	yes	1012GL	02-04-2022	no	
45	yes		21-01-2020	no	
51	no	8043VD	03-09-2009	yes	29-05-1995
45	no	1112XJ	12-01-2020	yes	13-04-2019
33	no	8011PK	15-04-2018	no	14-10-2015
22	yes	3572TC	27-08-2019	no	
29	no	3522BB	18-01-2013	yes	09-05-2010

Make use of ‘partially identifying variables’ *postal code, birth date*

Cluster records according to *non-linked from $\mathcal{A}$, non-linked from $\mathcal{B}$, linked common to $\mathcal{A}$ and $\mathcal{B}$*

The true linkage structure is latent



Record Linkage recipe

Record Linkage methods have been developed since the middle of the 20th century

Record Linkage recipe

Record Linkage methods have been developed since the middle of the 20th century

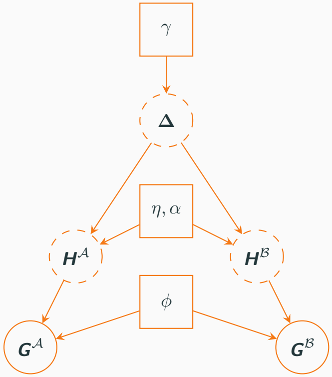
The old standard consists of a mixture model on the binary comparison of the records information

Record Linkage recipe

Record Linkage methods have been developed since the middle of the 20th century

The old standard consists of a mixture model on the binary comparison of the records information

New methodologies model the data generation process, taking account of registration errors



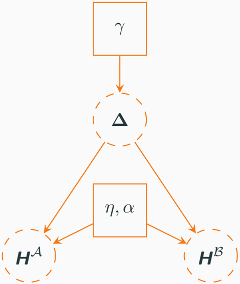
$\boldsymbol{G}^{\mathcal{A}}, \boldsymbol{G}^{\mathcal{B}}$ the registered values
$\boldsymbol{H}^{\mathcal{A}}, \boldsymbol{H}^{\mathcal{B}}$ the latent true values, $\boldsymbol{\Delta}$ the latent linkage matrix

FlexRL method

FlexRL uses a Stochastic EM approach to record linkage, (Robach et al., 2024)

It accounts for partially identifying variables that evolve through time (e.g. postal code) and handles large data sets

Teaser: we also develop a method to estimate the FDR in record linkage



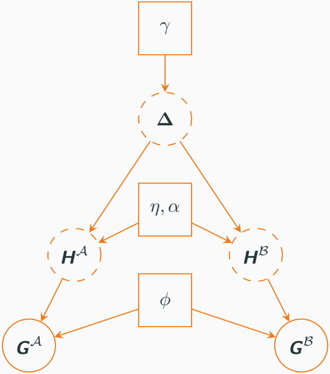
$$\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{G}^{\mathcal{A}}, \boldsymbol{G}^{\mathcal{B}}, \boldsymbol{H}^{\mathcal{A}}, \boldsymbol{H}^{\mathcal{B}}, \boldsymbol{\Delta}) = \mathcal{L}_{\phi}(\boldsymbol{G}^{\mathcal{A}} | \boldsymbol{H}^{\mathcal{A}}) \times \mathcal{L}_{\phi}(\boldsymbol{G}^{\mathcal{B}} | \boldsymbol{H}^{\mathcal{B}}) \\ \times \mathcal{L}_{\eta}(\boldsymbol{H}^{\mathcal{A}}) \times \mathcal{L}_{\alpha}(\boldsymbol{H}^{\mathcal{B}} | \boldsymbol{H}^{\mathcal{A}}, \boldsymbol{\Delta}) \times \mathcal{L}_{\gamma}(\boldsymbol{\Delta})$$

FlexRL method

FlexRL uses a Stochastic EM approach to record linkage, (Robach et al., 2024)

It accounts for partially identifying variables that evolve through time (e.g. postal code) and handles large data sets

Teaser: we also develop a method to estimate the FDR in record linkage



$$\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{G}^{\mathcal{A}}, \boldsymbol{G}^{\mathcal{B}}, \boldsymbol{H}^{\mathcal{A}}, \boldsymbol{H}^{\mathcal{B}}, \boldsymbol{\Delta}) = \mathcal{L}_{\phi}(\boldsymbol{G}^{\mathcal{A}} | \boldsymbol{H}^{\mathcal{A}}) \times \mathcal{L}_{\phi}(\boldsymbol{G}^{\mathcal{B}} | \boldsymbol{H}^{\mathcal{B}}) \\ \times \mathcal{L}_{\eta}(\boldsymbol{H}^{\mathcal{A}}) \times \mathcal{L}_{\alpha}(\boldsymbol{H}^{\mathcal{B}} | \boldsymbol{H}^{\mathcal{A}}, \boldsymbol{\Delta}) \times \mathcal{L}_{\gamma}(\boldsymbol{\Delta})$$

Stochastic EM for Record Linkage

A latent data problem

MLE $\hat{\boldsymbol{\theta}}_{ML}$ maximises

$$\sum_{\text{records}} \log \mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{G}^{\mathcal{A}}, \boldsymbol{G}^{\mathcal{B}})$$
$$= \sum_{\text{records}} \log \sum_{\boldsymbol{H}^{\mathcal{A}}} \sum_{\boldsymbol{H}^{\mathcal{B}}} \sum_{\boldsymbol{\Delta}} \mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{G}^{\mathcal{A}}, \boldsymbol{G}^{\mathcal{B}}, \boldsymbol{H}^{\mathcal{A}}, \boldsymbol{H}^{\mathcal{B}}, \boldsymbol{\Delta})$$

A latent data problem

MLE $\hat{\boldsymbol{\theta}}_{ML}$ maximises

$$\sum_{\text{records}} \log \mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{G}^{\mathcal{A}}, \boldsymbol{G}^{\mathcal{B}})$$

$$= \sum_{\text{records}} \log \sum_{\boldsymbol{H}^{\mathcal{A}}} \sum_{\boldsymbol{H}^{\mathcal{B}}} \sum_{\boldsymbol{\Delta}} \mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{G}^{\mathcal{A}}, \boldsymbol{G}^{\mathcal{B}}, \boldsymbol{H}^{\mathcal{A}}, \boldsymbol{H}^{\mathcal{B}}, \boldsymbol{\Delta})$$

StE-step $\rightarrow$ use a Gibbs sampler to generate true latent values $\boldsymbol{H}^{\mathcal{A}}, \boldsymbol{H}^{\mathcal{B}}$ of the partially identifying information and, the associated $\boldsymbol{\Delta}$

A latent data problem

MLE $\hat{\boldsymbol{\theta}}_{ML}$ maximises

$$\sum_{\text{records}} \log \mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{G}^{\mathcal{A}}, \boldsymbol{G}^{\mathcal{B}})$$
$$= \sum_{\text{records}} \log \sum_{\boldsymbol{H}^{\mathcal{A}}} \sum_{\boldsymbol{H}^{\mathcal{B}}} \sum_{\boldsymbol{\Delta}} \mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{G}^{\mathcal{A}}, \boldsymbol{G}^{\mathcal{B}}, \boldsymbol{H}^{\mathcal{A}}, \boldsymbol{H}^{\mathcal{B}}, \boldsymbol{\Delta})$$

StE-step $\rightarrow$ use a Gibbs sampler to generate true latent values $\boldsymbol{H}^{\mathcal{A}}, \boldsymbol{H}^{\mathcal{B}}$ of the partially identifying information and, the associated $\boldsymbol{\Delta}$

M-step $\rightarrow$ maximise the 'augmented' data log-likelihood and update the model parameters $\gamma, \eta, \alpha, \phi$

FlexRL model: an illustration

$\mathcal{A}$

zipcode	delivery date
1012GL	28-06-2021
1112XJ	18-04-2019
8043VD	14-10-2015
3572TC	03-08-2008

$\mathcal{B}$

zipcode	past delivery
1012GL	
8043VD	29-05-1995
1112XJ	13-04-2019
8011PK	14-10-2015
3572TC	
3522BB	09-05-2010

FlexRL model: an illustration

$\mathcal{A}$

zipcode	delivery date
1012GL	28-06-2021
1112XJ	18-04-2019
8043VD	14-10-2015
3572TC	03-08-2008

$\mathcal{B}$

zipcode	past delivery
1012GL	
8043VD	29-05-1995
1112XJ	13-04-2019
8011PK	14-10-2015
3572TC	
3522BB	09-05-2010

ϕ^t proportion of missing values and probability of mistakes in registered data

FlexRL model: an illustration

$$\mathcal{L}_{\phi^t}(\boldsymbol{G}^{\mathcal{A}} | \boldsymbol{H}^{\mathcal{A}}) \times \mathcal{L}_{\phi^t}(\boldsymbol{G}^{\mathcal{B}} | \boldsymbol{H}^{\mathcal{B}}) \times$$

$\mathcal{A}$

zipcode	delivery date
1012GL	28-06-2021
1112XJ	~~18-04-2019~~
8043VD	14-10-2015
3572TC	03-08-2008

$\mathcal{B}$

zipcode	past delivery
1012GL	?
?	?
8043VD	29-05-1995
1112XJ	13-04-2019
8011PK	14-10-2015
3572TC	?
~~3522BB~~	09-05-2010

ϕ^t proportion of missing values and probability of mistakes in registered data

FlexRL model: an illustration

$$\mathcal{L}_{\phi^t}(\boldsymbol{G}^{\mathcal{A}} | \boldsymbol{H}^{\mathcal{A}}) \times \mathcal{L}_{\phi^t}(\boldsymbol{G}^{\mathcal{B}} | \boldsymbol{H}^{\mathcal{B}}) \times$$

$\mathcal{A}$

zipcode	delivery date
1012GL	28-06-2021
1112XJ	~~18-04-2019~~
8043VD	14-10-2015
3572TC	03-08-2008

$\mathcal{B}$

zipcode	past delivery
1012GL	?
?	?
8043VD	29-05-1995
1112XJ	13-04-2019
8011PK	14-10-2015
3572TC	?
~~3522BB~~	09-05-2010

ϕ^t proportion of missing values and probability of mistakes in registered data

η^t distribution of the partially identifying variables

α^t probability of changes in information through time

FlexRL model: an illustration

$$\mathcal{L}_{\phi^t}(\boldsymbol{G}^{\mathcal{A}}|\boldsymbol{H}^{\mathcal{A}}) \times \mathcal{L}_{\phi^t}(\boldsymbol{G}^{\mathcal{B}}|\boldsymbol{H}^{\mathcal{B}}) \times \mathcal{L}_{\eta^t}(\boldsymbol{H}^{\mathcal{A}}) \times \mathcal{L}_{\alpha^t}(\boldsymbol{H}^{\mathcal{B}}|\boldsymbol{H}^{\mathcal{A}}, \boldsymbol{\Delta})$$

$\mathcal{A}$

zipcode	delivery date
1012GL	28-06-2021
1112XJ	13-04-2019
8043VD	14-10-2015
3572TC	03-08-2008

$\mathcal{B}$

zipcode	past delivery
1012GL	01-02-2003
1105AT	28-09-2006
8043VD	29-05-1995
1112XJ	13-04-2019
8011PK	14-10-2015
3572TC	08-12-2011
3526WP	09-05-2010

ϕ^t proportion of missing values and probability of mistakes in registered data

η^t distribution of the partially identifying variables

α^t probability of changes in information through time

FlexRL model: an illustration

$$\mathcal{L}_{\phi^t}(\boldsymbol{G}^{\mathcal{A}}|\boldsymbol{H}^{\mathcal{A}}) \times \mathcal{L}_{\phi^t}(\boldsymbol{G}^{\mathcal{B}}|\boldsymbol{H}^{\mathcal{B}}) \times \mathcal{L}_{\eta^t}(\boldsymbol{H}^{\mathcal{A}}) \times \mathcal{L}_{\alpha^t}(\boldsymbol{H}^{\mathcal{B}}|\boldsymbol{H}^{\mathcal{A}}, \boldsymbol{\Delta})$$

$\mathcal{A}$

zipcode	delivery date
1012GL	28-06-2021
1112XJ	13-04-2019
8043VD	14-10-2015
3572TC	03-08-2008

$\mathcal{B}$

zipcode	past delivery
1012GL	01-02-2003
1105AT	28-09-2006
8043VD	29-05-1995
1112XJ	13-04-2019
8011PK	14-10-2015
3572TC	08-12-2011
3526WP	09-05-2010

ϕ^t proportion of missing values and probability of mistakes in registered data

η^t distribution of the partially identifying variables

α^t probability of changes in information through time

γ^t proportion of links

FlexRL model: an illustration

$$\mathcal{L}_{\phi^t}(\boldsymbol{G}^{\mathcal{A}}|\boldsymbol{H}^{\mathcal{A}}) \times \mathcal{L}_{\phi^t}(\boldsymbol{G}^{\mathcal{B}}|\boldsymbol{H}^{\mathcal{B}}) \times \mathcal{L}_{\eta^t}(\boldsymbol{H}^{\mathcal{A}}) \times \mathcal{L}_{\alpha^t}(\boldsymbol{H}^{\mathcal{B}}|\boldsymbol{H}^{\mathcal{A}}, \boldsymbol{\Delta}) \times \mathcal{L}_{\gamma^t}(\boldsymbol{\Delta})$$

$\mathcal{A}$

zipcode	delivery date
1012GL	28-06-2021
1112XJ	13-04-2019
8043VD	14-10-2015
3572TC	03-08-2008

$\mathcal{B}$

zipcode	past delivery
1012GL	01-02-2003
1105AT	28-09-2006
8043VD	29-05-1995
1112XJ	13-04-2019
8011PK	14-10-2015
3572TC	08-12-2011
3526WP	09-05-2010

- ϕ^t proportion of missing values and probability of mistakes in registered data
- η^t distribution of the partially identifying variables
- α^t probability of changes in information through time
- γ^t proportion of links

FlexRL model: an illustration

$$\mathcal{L}_{\phi^t}(\boldsymbol{G}^{\mathcal{A}}|\boldsymbol{H}^{\mathcal{A}}) \times \mathcal{L}_{\phi^t}(\boldsymbol{G}^{\mathcal{B}}|\boldsymbol{H}^{\mathcal{B}}) \times \mathcal{L}_{\eta^t}(\boldsymbol{H}^{\mathcal{A}}) \times \mathcal{L}_{\alpha^t}(\boldsymbol{H}^{\mathcal{B}}|\boldsymbol{H}^{\mathcal{A}}, \boldsymbol{\Delta}) \times \mathcal{L}_{\gamma^t}(\boldsymbol{\Delta})$$

$\mathcal{A}$

zipcode	delivery date
1012GL	28-06-2021
1112XJ	13-04-2019
8043VD	14-10-2015
3572TC	03-08-2008

$\mathcal{B}$

zipcode	past delivery
1012GL	01-02-2003
1105AT	28-09-2006
8043VD	29-05-1995
1112XJ	13-04-2019
8011PK	14-10-2015
3572TC	08-12-2011
3526WP	09-05-2010

ϕ^{t+1} proportion of missing values and probability of mistakes in registered data

η^{t+1} distribution of the partially identifying variables

α^{t+1} probability of changes in information through time

γ^{t+1} proportion of links

Real data applications

Perinatal registry: Utrecht province

Pregnancy data from 1999 until 2009 for the province of Utrecht in the Netherlands

We link the 1st born (7000 records) and 2nd born children (1500 records)

PIVs	Unique values	Type	Missing	Agreements in true links
Postal Code	25	categorical	0	?
M dob yy	30	categorical	0	?
M dob mm	12	categorical	0	?
M dob dd	31	categorical	0	?
S dob yy	11	categorical	0	?
S dob mm	12	categorical	0	?
S dob dd	31	categorical	0	?

We have no unique identifier on these data sets (we validated the method on common data sets from the literature with unique identifiers)

Results

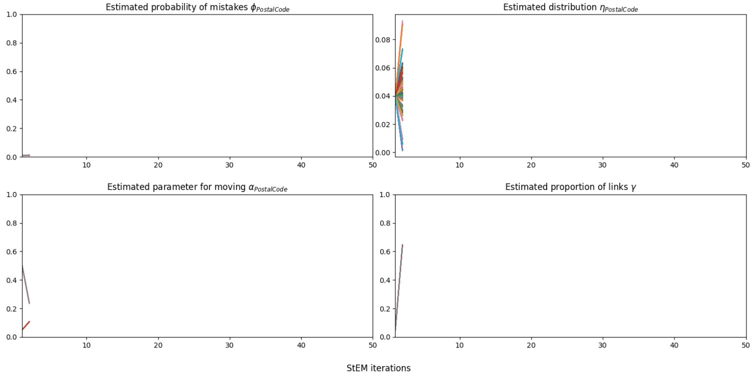
- simplistic: links records with matching information
- *BRL*: enhances the foundational mixture model (Sadinle, 2017)
- *FastLink*: fast and scalable version of *BRL* (Enamorado et al., 2019)

Methods	Simplistic	*FlexRL all stable*	*FlexRL with dynamic Postal Code*	*BRL*	*FastLink*
Linked records	898	889	988	1005	1006
Estimated FDR	.01	.00	.00	.01	.01

Agreements	Simplistic	*FlexRL all stable*	*FlexRL with dynamic Postal Code*	*BRL*	*FastLink*
Postal Code	1	.99	.94	.94	.94
M dob yy	1	1	.99	.99	.99
M dob mm	1	1	.99	.99	.99
M dob dd	1	1	.99	.99	.99
S dob yy	1	1	.99	.99	.99
S dob mm	1	1	.99	.99	.99
S dob dd	1	1	.99	.97	.97

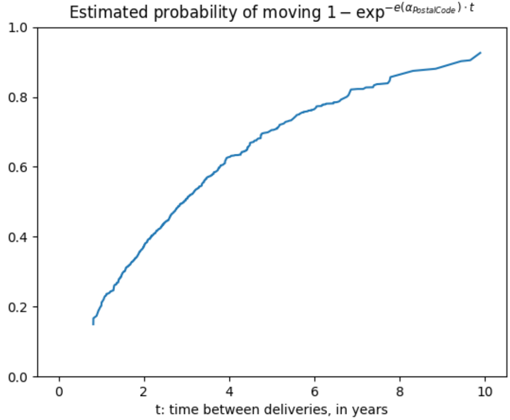
Convergence of ***FlexRL***

Probability of mistakes, distribution of the PIV, changes across time, for *Postal Code* and, proportion of linked records



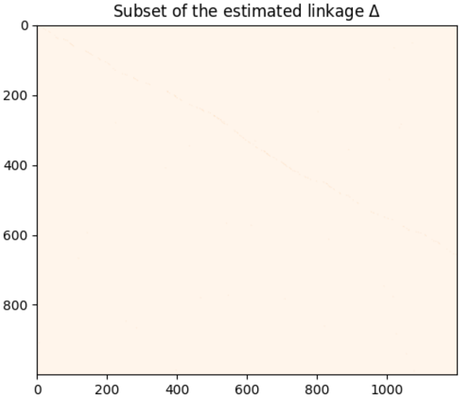
Convergence of ***FlexRL***

Estimated probability of moving between the deliveries in the province of Utrecht



Convergence of *FlexRL*

Estimated linkage matrix between 1st and 2nd born children



NLTCS registry

Data from a longitudinal survey on edlery care in the US (1982 and 1994) with 20500 and 9500 records

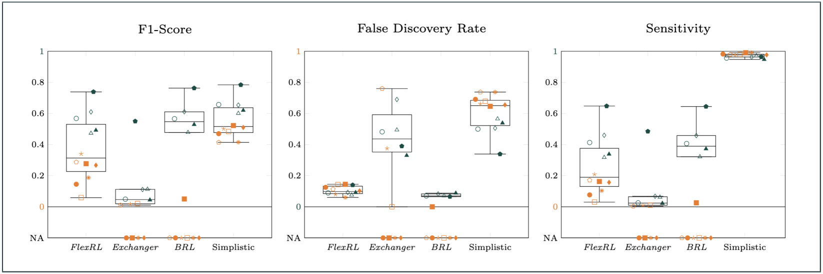
Registrations		Sex	Birth month	Birth year	State code	Regional code
Data	Unique	2	12	57	58	12
	Type	categorical	categorical	categorical	categorical	categorical
	Missing	0	0	0	0	.02
True Links	Agree	1	1	1	.91	.92

Characteristics of the PIVs and level of agreement among the 7500 links referring to the same individuals

We have a unique identifier to validate the method

NLTCS registry: regional applications

- simplistic: links records with matching information
- *BRL*: enhances the foundational mixture model (Sadinle, 2017)
- *Exchanger*: graphical entity resolution model (Marchant et al., 2023)



References

Celeux, G. and Diebolt, J. (1986). L'algorithme SEM : un algorithme d'apprentissage probabiliste pour la reconnaissance de mélange de densités. *Revue de Statistiques Appliquées*, 34(2):35–52.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1).

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Enamorado, T., Fifield, B., and Imai, K. (2019). Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, 113(2):353–371.

Marchant, N. G., Rubinstein, B. I. P., and Steorts, R. C. (2023). Bayesian Graphical Entity Resolution using Exchangeable Random Partition Priors. *Journal of Survey Statistics and Methodology*, 11(3):569–596.

Robach, K., van der Pas, S., van de Wiel, M., and Hof, M. H. (2024). A flexible model for record linkage.

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

Wu, C. F. J. (1983). On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95 – 103.

Thank You!

Appendix

SHIW

Data from a longitudinal survey of Household Income and Wealth in Italy (2016 and 2020) with 15000 and 16500 records

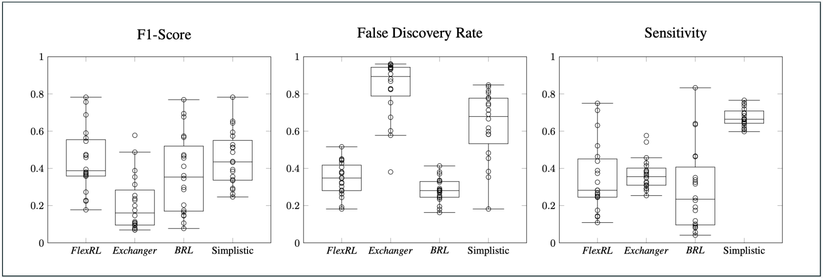
Registrations		Sex	Birth year	Marital status	Regional code	Birth region	Education
Data	Unique	2	97	4	20	20	6
	Type	categorical	categorical	categorical	categorical	categorical	categorical
	Missing	0	0	0	0	.05	0
True Links	Agree	1	.98	.94	1	.94	.77

Characteristics of the PIVs and level of agreement among the 6400 links referring to the same individuals

We have a unique identifier to validate the method

Results on regional subsets

- simplistic: links records with matching information
- *BRL*: enhances the foundational mixture model (Sadinle, 2017)
- *Exchanger*: graphical entity resolution model (Marchant et al., 2023)



FDR estimation

Estimating the FDR is important to apply record linkage on real use cases

$\text{FDR}(\xi) = \mathbb{E}\left[\frac{FP(\xi)}{TP(\xi)+FP(\xi)}\right]$ depends on a ξ to define what is 'positive'

FDR estimation

Estimating the FDR is important to apply record linkage on real use cases

$$\mathrm{FDR}(\xi) = \mathbb{E}\left[\frac{FP(\xi)}{TP(\xi)+FP(\xi)}\right]$$

Recipe:

- Synthesise data

FDR estimation

Estimating the FDR is important to apply record linkage on real use cases

$$\text{FDR}(\xi) = \mathbb{E}\left[\frac{FP(\xi)}{TP(\xi)+FP(\xi)}\right]$$

Recipe:

- Synthesise data with the same underlying structure (conditional distributions fitted to the original data)

FDR estimation

Estimating the FDR is important to apply record linkage on real use cases

$$\mathrm{FDR}(\xi) = \mathbb{E}\left[\frac{FP(\xi)}{TP(\xi)+FP(\xi)}\right]$$

Recipe:

- Synthesise data with the same underlying structure (conditional distributions fitted to the original data); similar proportions of FP in real and synthetic data are expected

FDR estimation

Estimating the FDR is important to apply record linkage on real use cases

$$\text{FDR}(\xi) = \mathbb{E}\left[\frac{FP(\xi)}{TP(\xi)+FP(\xi)}\right]$$

Recipe:

- Synthesise data
  - linked pairs involving a synthetic records are $FP_{\text{synthetic}}(\xi)$
  - the total number of linked pairs is $FP_{\text{synthetic}}(\xi) + \underbrace{FP(\xi) + TP(\xi)}_{\text{real linked pairs}}$

FDR estimation

Estimating the FDR is important to apply record linkage on real use cases

$$\mathrm{FDR}(\xi) = \mathbb{E}\left[\frac{FP(\xi)}{TP(\xi)+FP(\xi)}\right]$$

Recipe:

- Synthesise data
  - linked pairs involving a synthetic records are $FP_{\text{synthetic}}(\xi)$
  - the total number of linked pairs is $FP_{\text{synthetic}}(\xi) + \underbrace{FP(\xi) + TP(\xi)}_{\text{real linked pairs}}$
- $\widehat{\mathrm{FDR}}(\xi) = \frac{FP_{\text{synthetic}}(\xi)}{\text{real linked pairs}(\xi)}$

FDR estimation

Estimating the FDR is important to apply record linkage on real use cases

$$\text{FDR}(\xi) = \mathbb{E}\left[\frac{FP(\xi)}{TP(\xi)+FP(\xi)}\right]$$

Recipe:

- Synthesise data
  - linked pairs involving a synthetic records are $FP_{\text{synthetic}}(\xi)$
  - the total number of linked pairs is $FP_{\text{synthetic}}(\xi) + \underbrace{FP(\xi) + TP(\xi)}_{\text{real linked pairs}}$
- $\widehat{\text{FDR}}(\xi) = \frac{FP_{\text{synthetic}}(\xi)}{\text{real linked pairs}(\xi)}$ or, $\widehat{\text{FDR}}(\xi) = \frac{2 \cdot FP_{\text{synthetic}}(\xi)}{\text{linked pairs}(\xi)}$

FDR estimation

Estimating the FDR is important to apply record linkage on real use cases

$$\text{FDR}(\xi) = \mathbb{E}\left[\frac{FP(\xi)}{TP(\xi)+FP(\xi)}\right]$$

Recipe:

- Synthesise data
  - linked pairs involving a synthetic records are $FP_{\text{synthetic}}(\xi)$
  - the total number of linked pairs is $FP_{\text{synthetic}}(\xi) + \underbrace{FP(\xi) + TP(\xi)}_{\text{real linked pairs}}$
- $\widehat{\text{FDR}}(\xi) = \frac{FP_{\text{synthetic}}(\xi)}{\text{real linked pairs}(\xi)}$ or, $\widehat{\text{FDR}}(\xi) = \frac{2 \cdot FP_{\text{synthetic}}(\xi)}{\text{linked pairs}(\xi)}$

The 2^{th} formula is more suited for large data sets applications