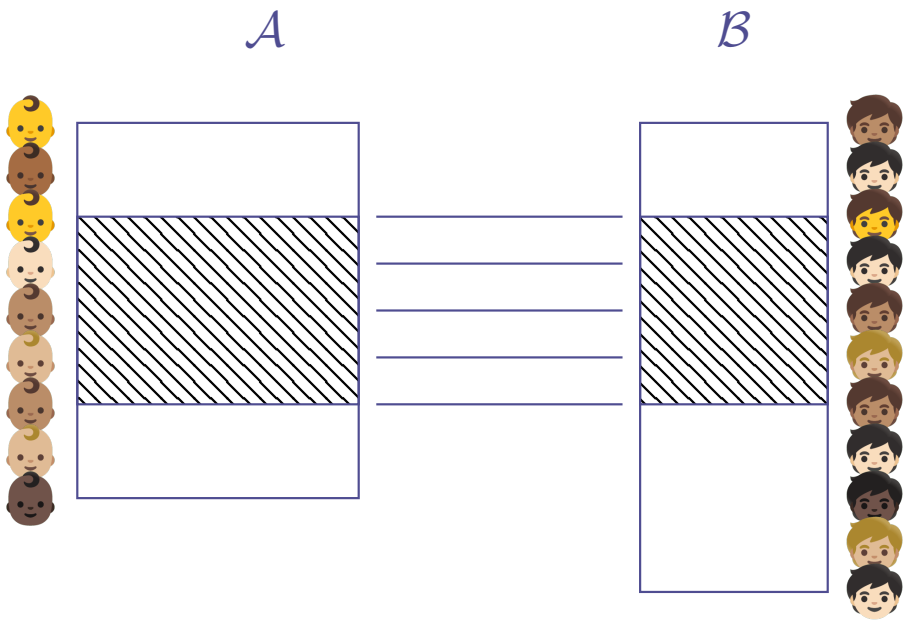
A Flexible Record Linkage Model

Kayané Robach

Joint work with Michel Hof, Stéphanie van der Pas, Mark van de Wiel

Case study

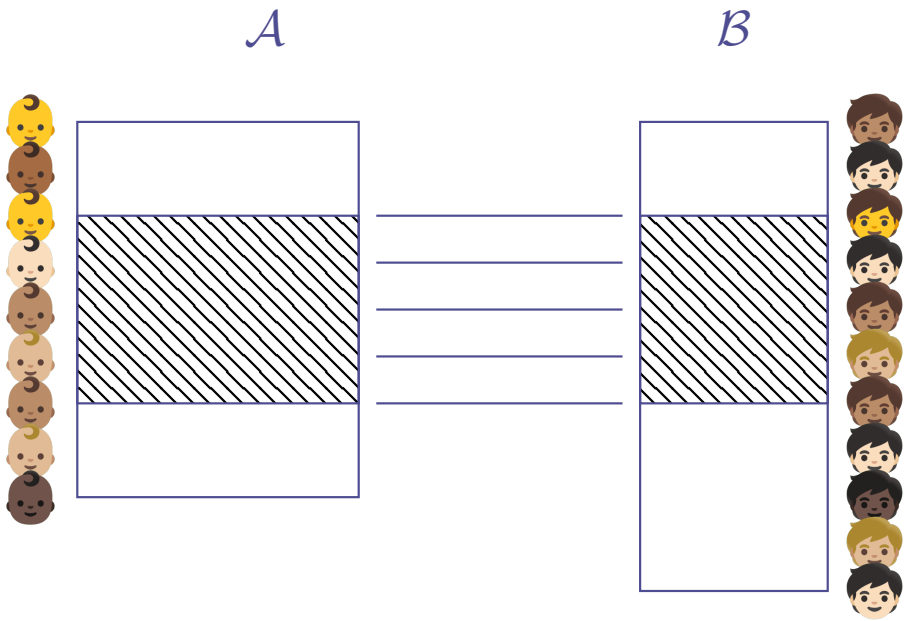
- How to merge multiple data sources? *Record Linkage*



Case study

Netherlands Perinatal Registry by Perined (PRN)

- Pregnancies and deliveries data about 96% of all births in the Netherlands



- Study in particular the risk of pre-term birth using data from past deliveries

🧮 Observations are the deliveries

$\Longrightarrow$ No unique identifier to assemble the family portraits

Case study

Toy example

- Make use of 'partially' identifying variables (common to the multiple sources)
  - place of residence (zipcode),
  - mother birth year,
  - date of delivery / date of previous delivery

$\mathcal{A}$

zipcode	birth year	delivery date	pre-term
1012GL	1998	28-06-2021	yes
1112XJ	1978	13-04-2019	no
8043VD	1990	14-10-2015	yes

$\mathcal{B}$

Occupation	age	zipcode	birth year	delivery date	past deliveries
Researcher	25	1012GL	1998	02-04-2022	
Dancer	45	1112XJ	1978	12-01-2020	13-04-2019
Baker	33	8011PK	1990	15-04-2018	14-10-2015
Teacher	45	1112XJ	1978	21-01-2020	
GP	51	3011CC	1972	03-09-2000	29-05-1995

Questioning the old standard

Popular method: compact the information into *comparison vectors*

- easy to handle as i.i.d. data from a mixture between linked and non-linked records
- ignores the weak dependencies among the comparison vectors
- thus treats linkage decisions independently from one another

$\rightarrow$ Raises the need for a post-hoc step to ensure a *one-to-one assignment*

$\rightarrow$ Independence among comparison vectors and linkage decisions is not valid

Fresh approaches

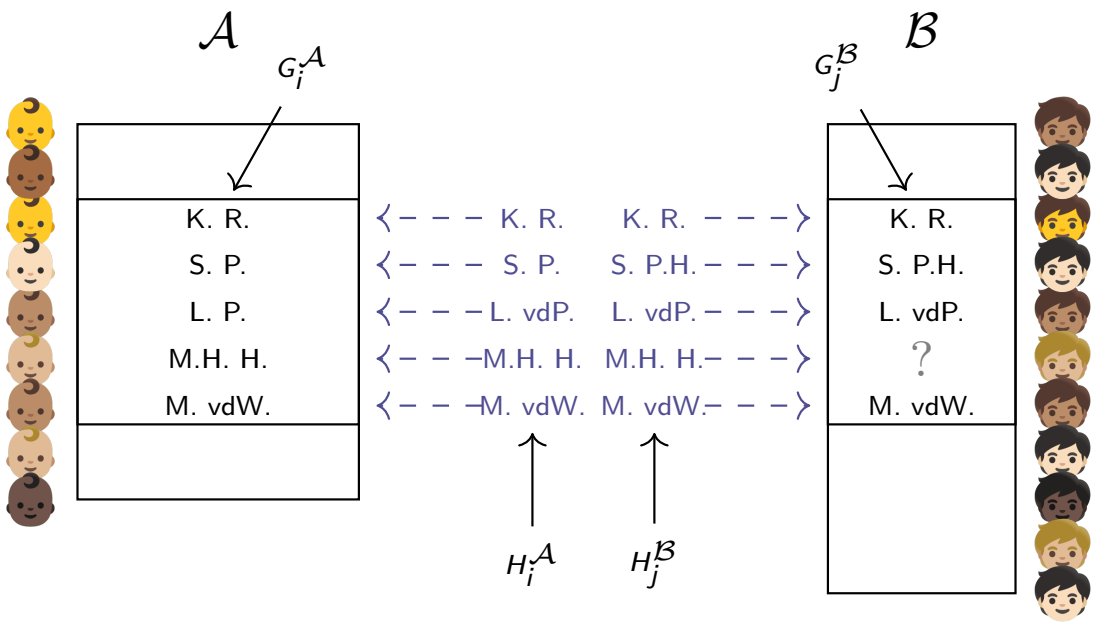
One solution suggests to use the complete data

- model the data generating process
- include distortion mechanisms explaining registration errors in the data

[Tancredi and Liseo, 2011] develop such method but it is computationally heavy, not scalable to big datasets.

[Steorts et al., 2015] also model the data generating process but still use the comparison vectors.

A short manual of record linkage



- Model the data generating process
  - no reduction of information
  - can incorporate any plausible distortion mechanisms
  - fewer possible linkages given true values $H_i^{\mathcal{A}}, H_j^{\mathcal{B}}$
- Consider dependent linkage decisions

A short manual of record linkage

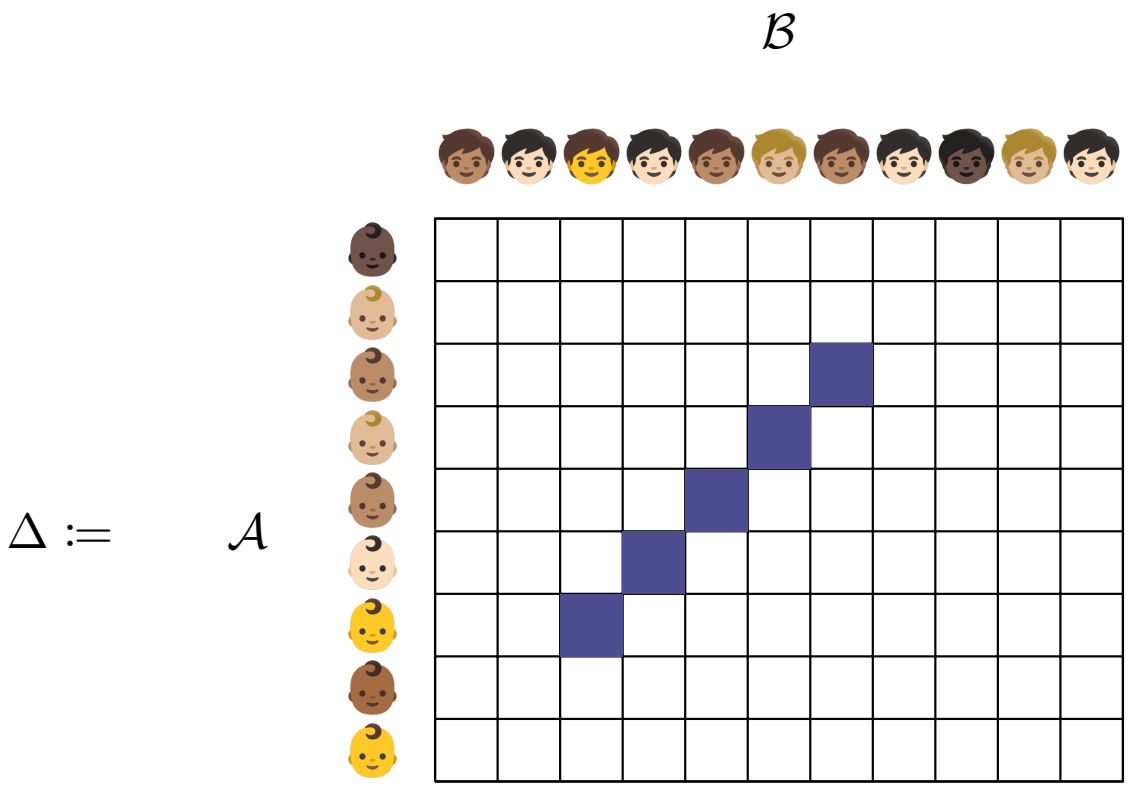
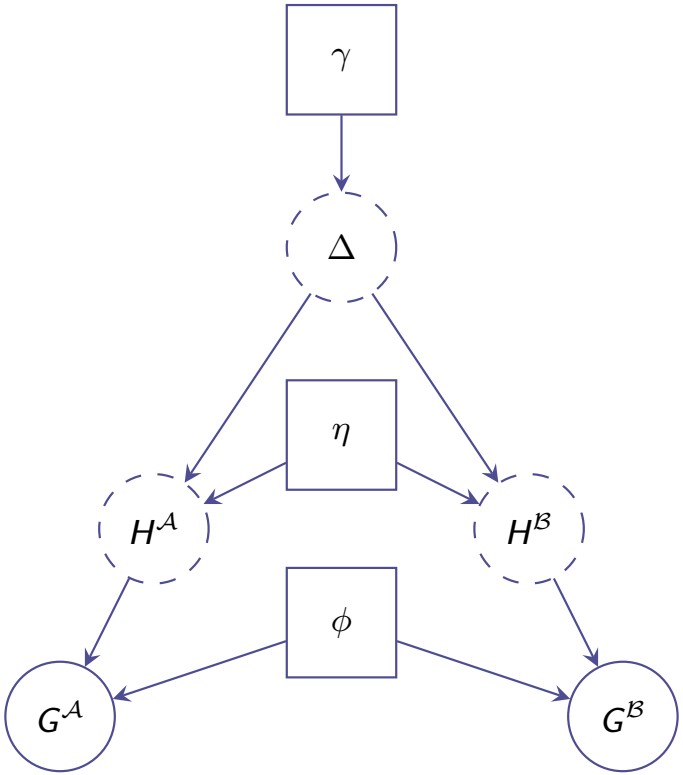
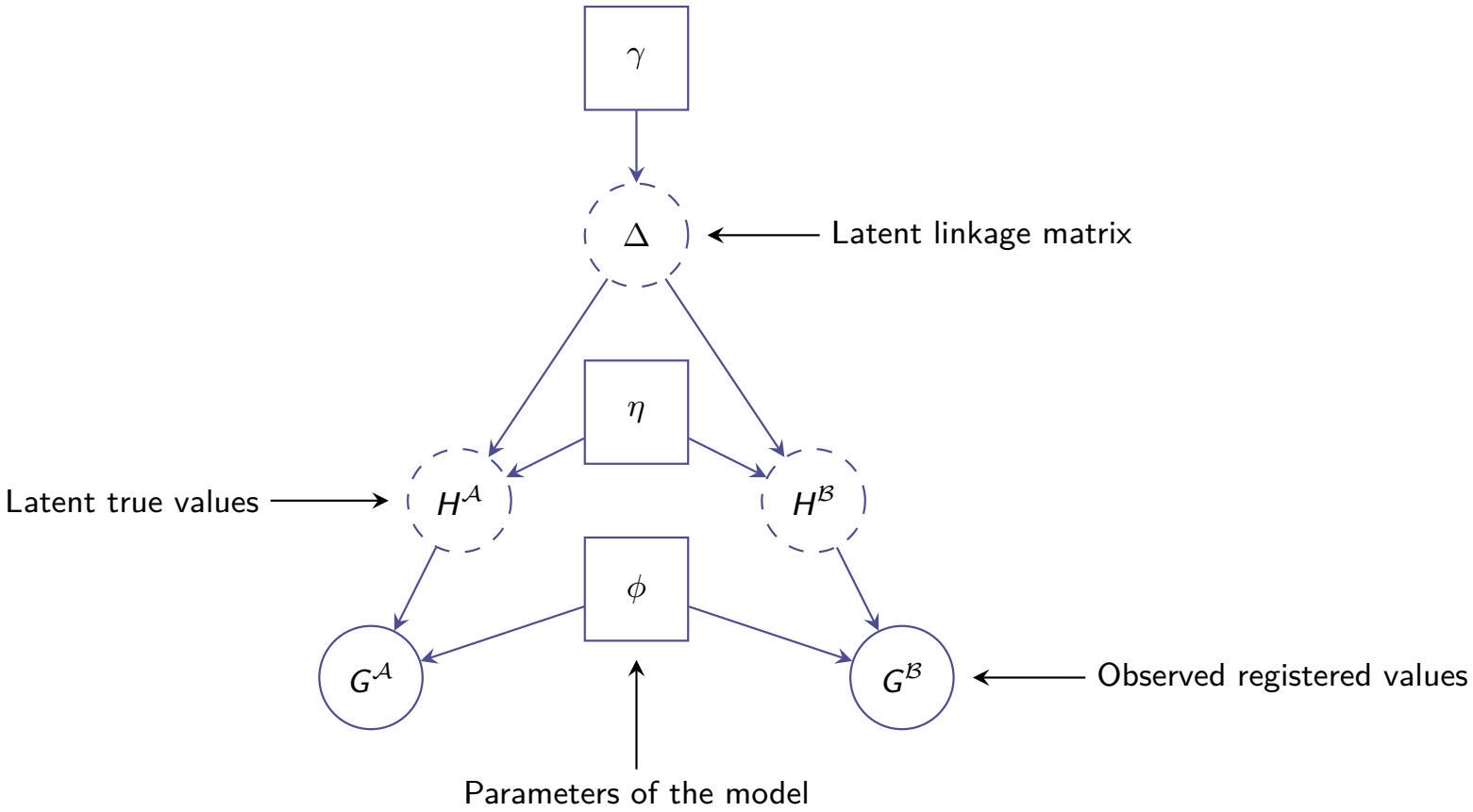


Figure: Illustration for the true linkage matrix Δ we would like to estimate.

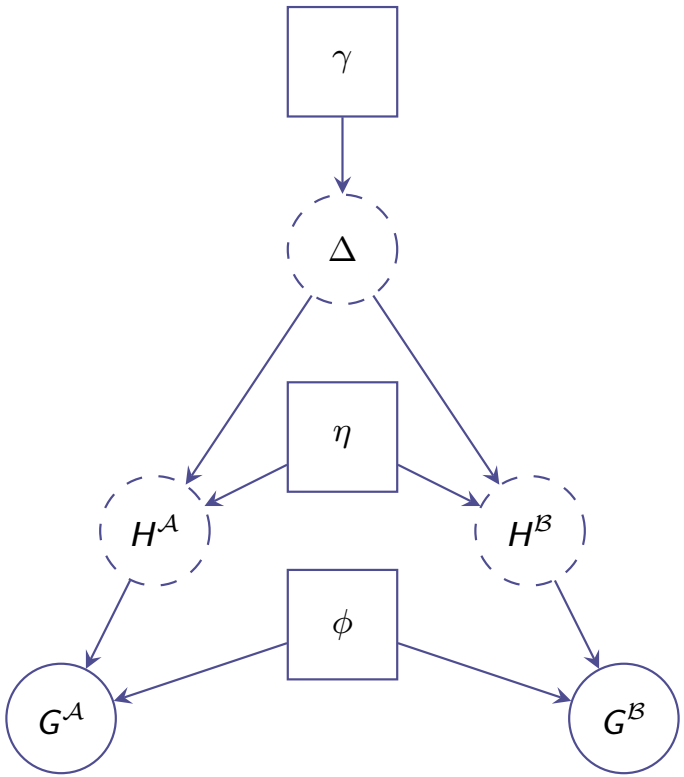
The probabilistic graph of our new model



The probabilistic graph of our new model

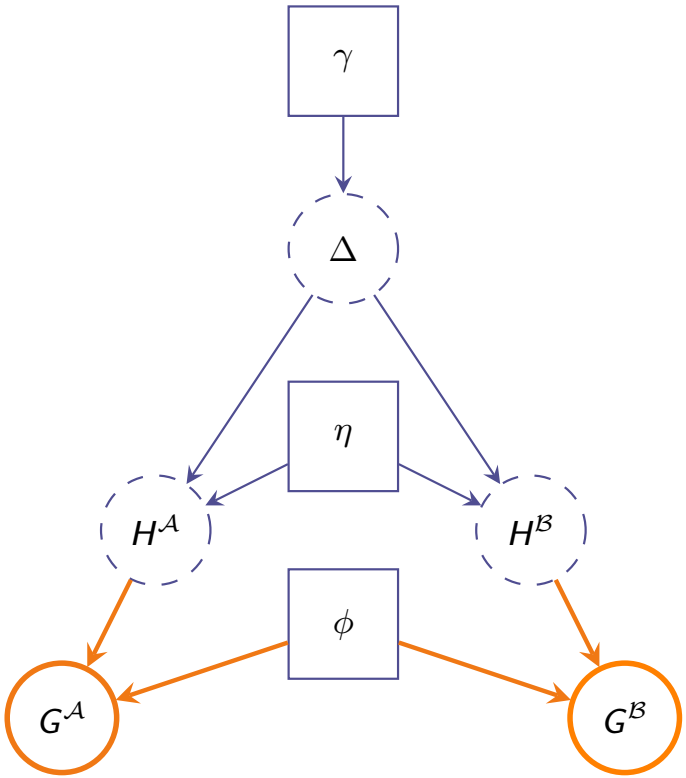


New model



We treat H and Δ as missing values and use a standard MC-EM algorithm.

Registered values



The registered partially identifying variables: $G^{\mathcal{A}}$, $G^{\mathcal{B}}$ and their model[a]

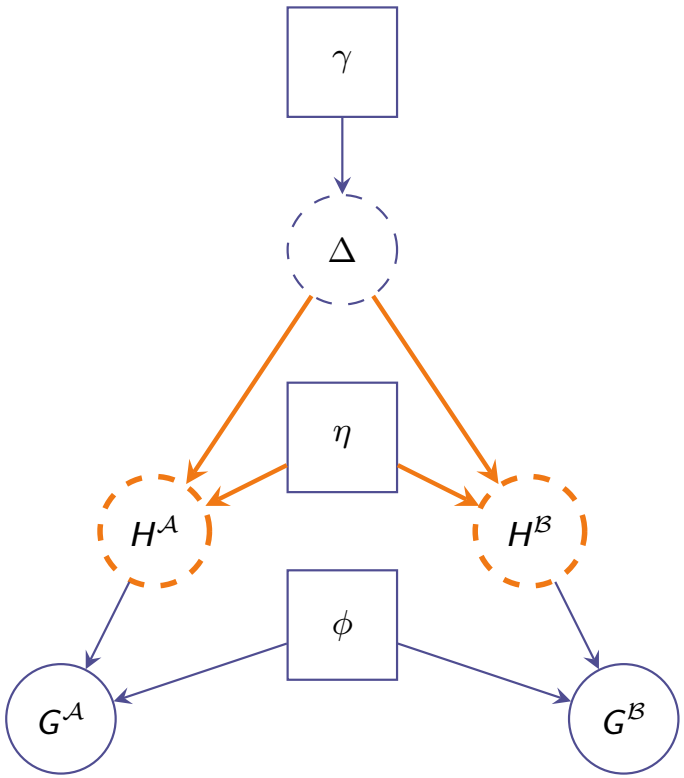
$$\mathbb{P}\left(G_i^{\mathcal{A}} = a \,\middle|\, H_i^{\mathcal{A}} = b; \phi\right) =$$
$$\mathbb{P}\left(G_j^{\mathcal{B}} = a \,\middle|\, H_j^{\mathcal{B}} = b; \phi\right)$$

where we can incorporate

- missing registered values,
- mistakes / typos in registered values (compared to the truth)

[a] Observations in $\mathcal{A}$ and in $\mathcal{B}$ comes from the same super population, hence the same distribution for both set of individuals.

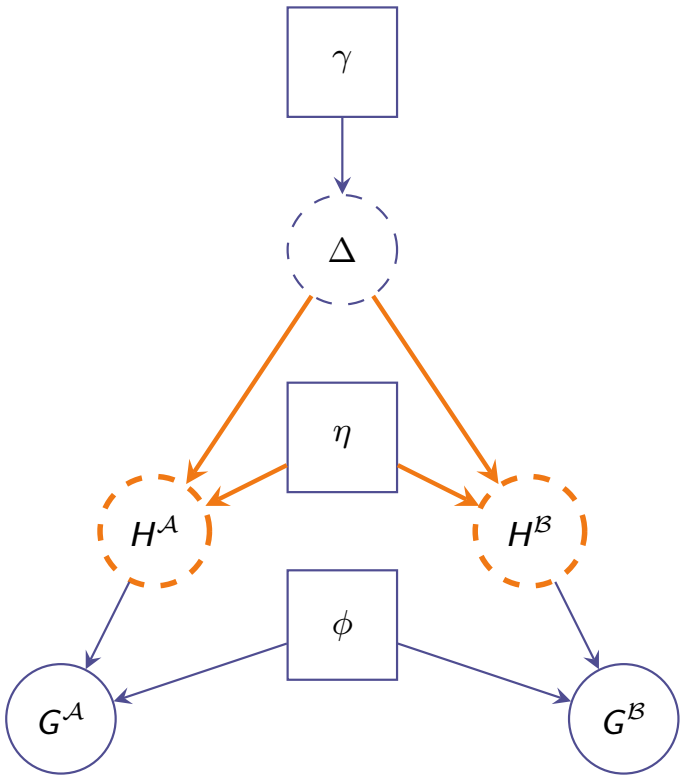
Latent true values



The latent partially identifying variables $H^{\mathcal{A}}$, $H^{\mathcal{B}}$ and

- a model for the values distribution $\mathbb{P}\left(H_i^{\mathcal{A}} = a; \eta\right) = \mathbb{P}\left(H_j^{\mathcal{B}} = a; \eta\right)$,
- a joint model for linked records $\mathbb{P}\left(H_i^{\mathcal{A}} = a, H_j^{\mathcal{B}} = b \,\middle|\, \Delta_{i,j} = 1; \eta\right)$

Latent true values

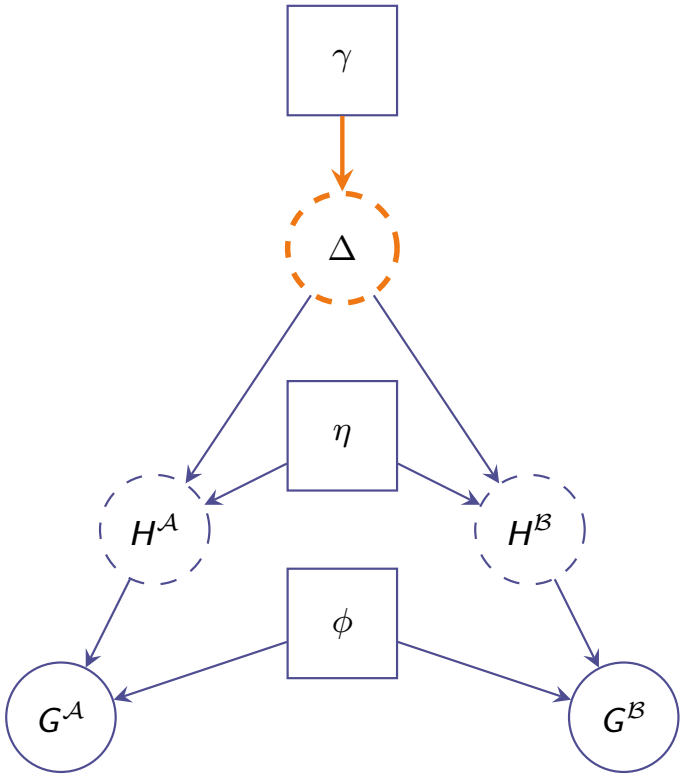


The latent partially identifying variables $H^{\mathcal{A}}$, $H^{\mathcal{B}}$ and

- a model for the values distribution $\mathbb{P}\left(H_i^{\mathcal{A}} = a; \eta\right) = \mathbb{P}\left(H_j^{\mathcal{B}} = a; \eta\right)$,
- a joint model for linked records $\mathbb{P}\left(H_i^{\mathcal{A}} = a; \eta\right) \cdot \mathbb{P}\left(H_i^{\mathcal{A}} = a \,\middle|\, H_j^{\mathcal{B}} = b, \Delta_{i,j} = 1; \eta\right)$

where we can incorporate unstable partially identifying variables that change over time

Linkage matrix



Our parameter of interest:

$$\Delta = \begin{pmatrix} \Delta_{1,1} & \Delta_{1,2} & \cdots & \Delta_{1,n_{\mathcal{B}}} \\ \Delta_{2,1} & \Delta_{2,2} & \cdots & \Delta_{2,n_{\mathcal{B}}} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta_{n_{\mathcal{A}},1} & \Delta_{n_{\mathcal{A}},2} & \cdots & \Delta_{n_{\mathcal{A}},n_{\mathcal{B}}} \end{pmatrix}$$

with its definition set:

$$D = \Bigg\{ \Delta : \Delta_{i,j} \in \{0,1\}, \sum_{i=1}^{n_{\mathcal{A}}} \Delta_{i,j} \leq 1 \text{ for all } j \in [\![1, n_{\mathcal{B}}]\!] \text{ and } \sum_{j=1}^{n_{\mathcal{B}}} \Delta_{i,j} \leq 1 \text{ for all } i \in [\![1, n_{\mathcal{A}}]\!] \Bigg\}$$

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

Our approach

→ **MC-EM** makes the process scalable to big datasets

→ Submodels can incorporate extra information to **suit other tasks**

→ **MLE** approach leads to unbiased estimates
+ enables to use likelihood ratios and compare efficiency with other methods

→ The method can handle **partially identifying variables that change over time**
for e.g. place of residence, marital status

Linkage matrix

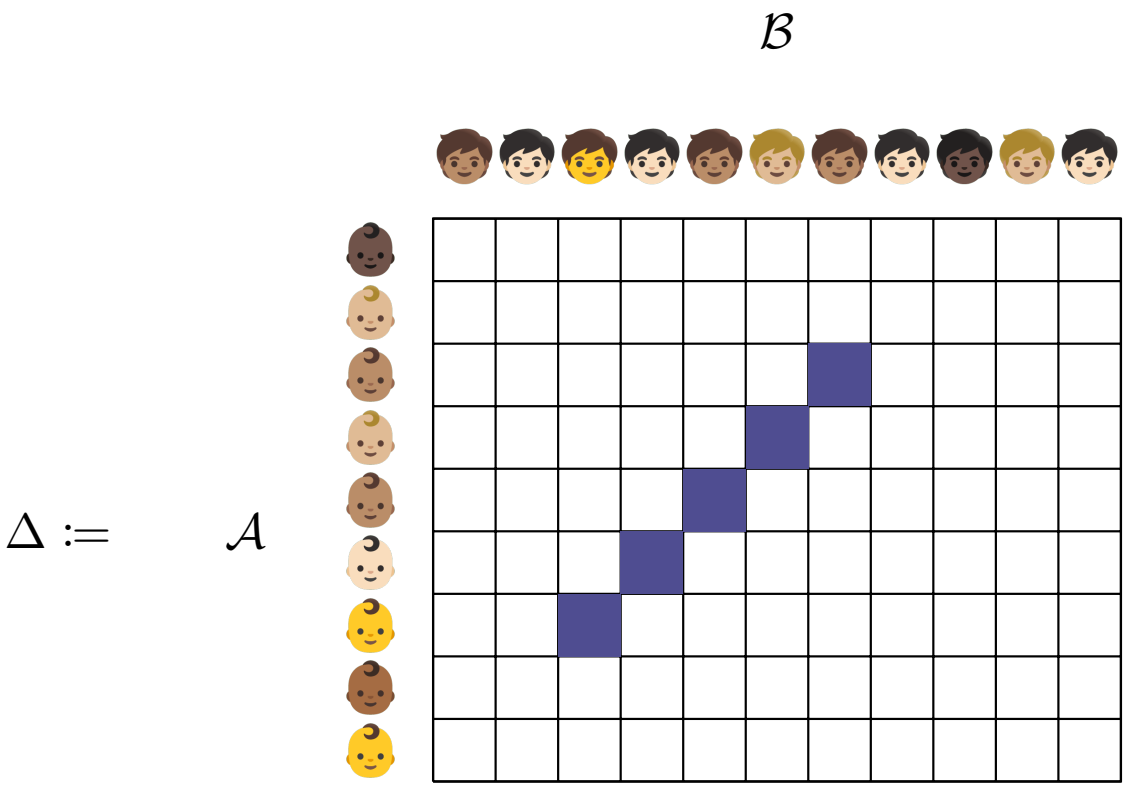


Figure: The linkage matrix Δ we would like to estimate.

Linkage matrix

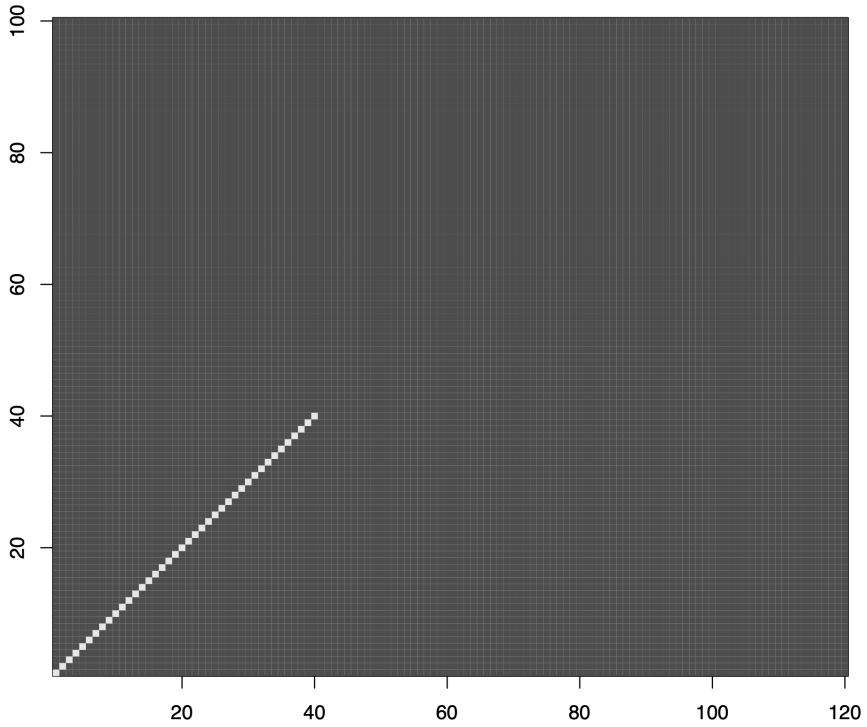


Figure: True Δ targeted in the simulation.

Linkage matrix

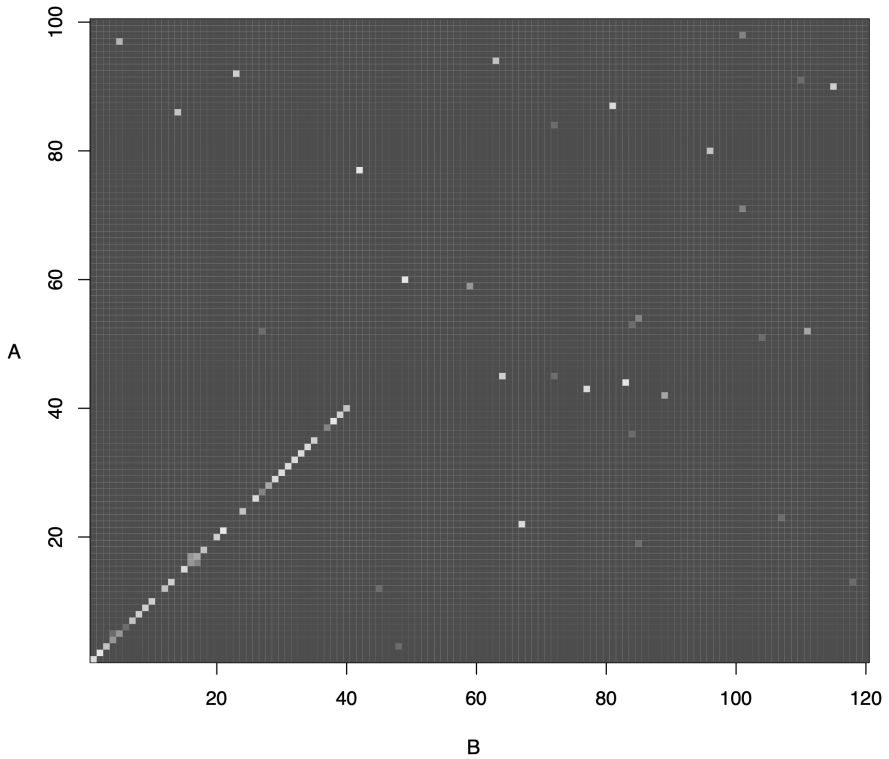


Figure: Result of the estimation of Δ.

Further work

The flexibility of the model allows us to

- propagate the uncertainty of the linkage to inference for causal questions
- add knowledge in $\mathbb{P}(\Delta)$ useful for inference

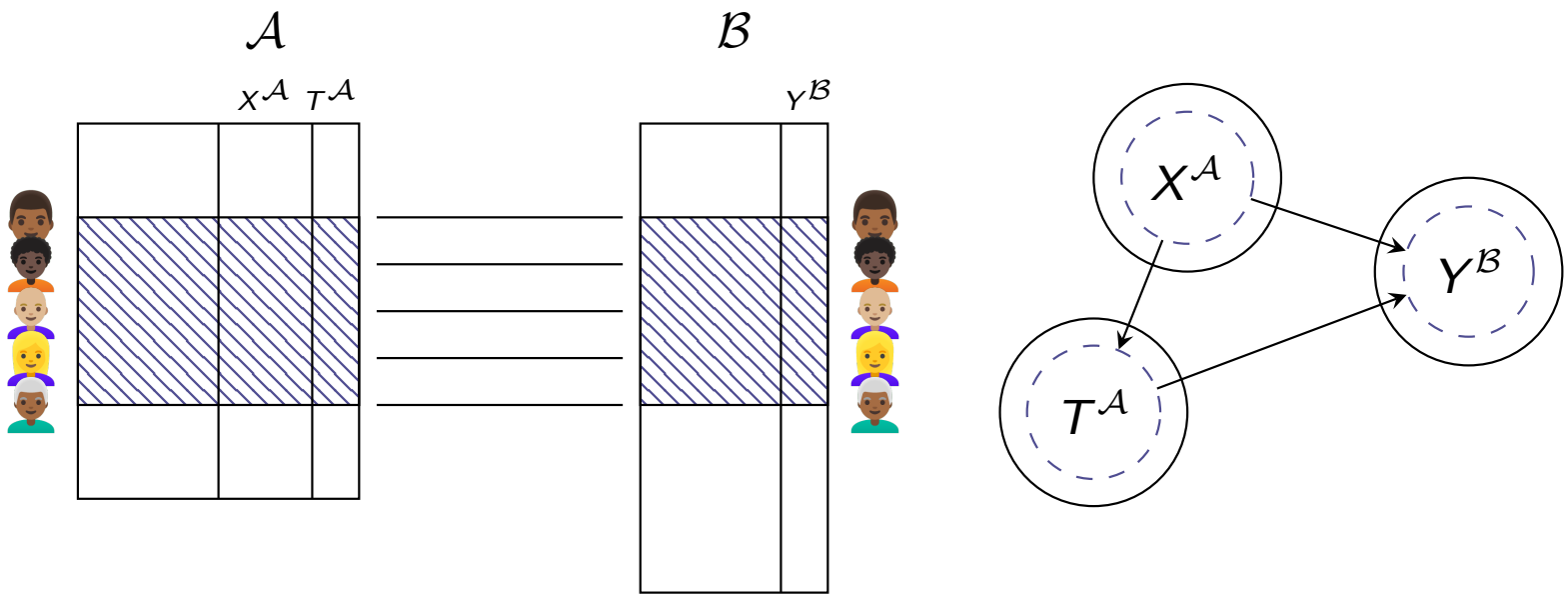


Figure: Illustration of a causality study with covariates X and treatment T in $\mathcal{A}$, outcomes Y in $\mathcal{B}$ for e.g. long-term outcomes studies, survival data

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

Thank You!

k.c.robach@amsterdamumc.nl

Steorts, R. C., Hall, R., and Fienberg, S. E. (2015).
A Bayesian approach to graphical record linkage and de-duplication.

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

Questioning the old standard

Independence among comparison vectors and linkage decisions is not valid

	mother birth year	zipcode
$1_{\mathcal{A}}, 1_{\mathcal{B}}$	0	1
$1_{\mathcal{A}}, 2_{\mathcal{B}}$	1	0
$2_{\mathcal{A}}, 1_{\mathcal{B}}$	1	0
$2_{\mathcal{A}}, 2_{\mathcal{B}}$	?	?

Figure: Example of comparison vectors built on $\mathcal{A} \times \mathcal{B}$ when both only contains two records: $1_{\mathcal{A}}, 2_{\mathcal{A}}$ and $1_{\mathcal{B}}, 2_{\mathcal{B}}$. For the mother's birth year we can easily compare the records, $2_{\mathcal{B}} = 1_{\mathcal{A}} \neq 1_{\mathcal{B}} = 2_{\mathcal{A}} \implies 2_{\mathcal{B}} \neq 2_{\mathcal{A}}$ by transitivity. For zipcode, this toy example is more complex, $2_{\mathcal{B}} \neq 1_{\mathcal{A}} = 1_{\mathcal{B}} \neq 2_{\mathcal{A}}$ and we do not know about the comparison between $2_{\mathcal{A}}$ and $2_{\mathcal{B}}$ in the end.

- we may deduce some values of the comparison vectors from the others,
- some fields may be correlated for e.g. day and month in dates across EU or US,
- the one-to-one assignment $\implies$ dependencies among linkage decisions

Simulations

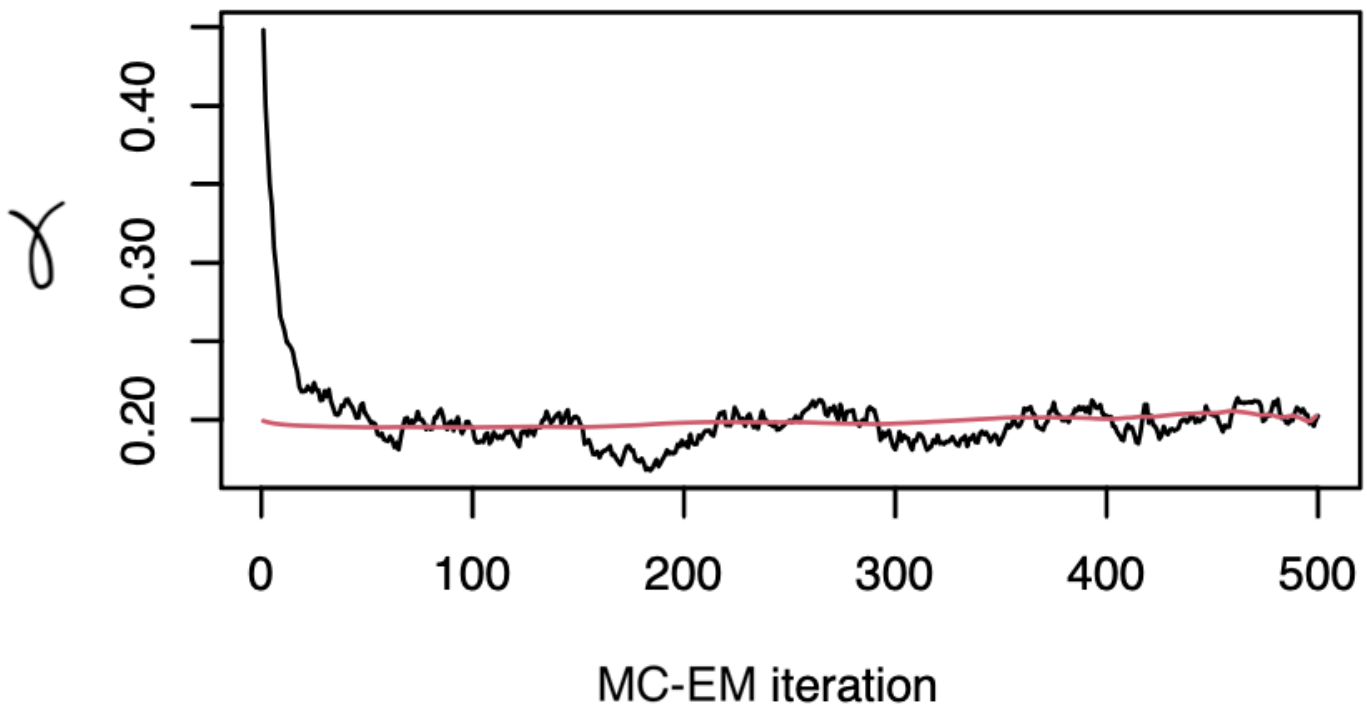


Figure: Convergence of the probability to have a link (truth: 20% of the smallest file) in an MC-EM algorithm with 500 iterations.

Simulations

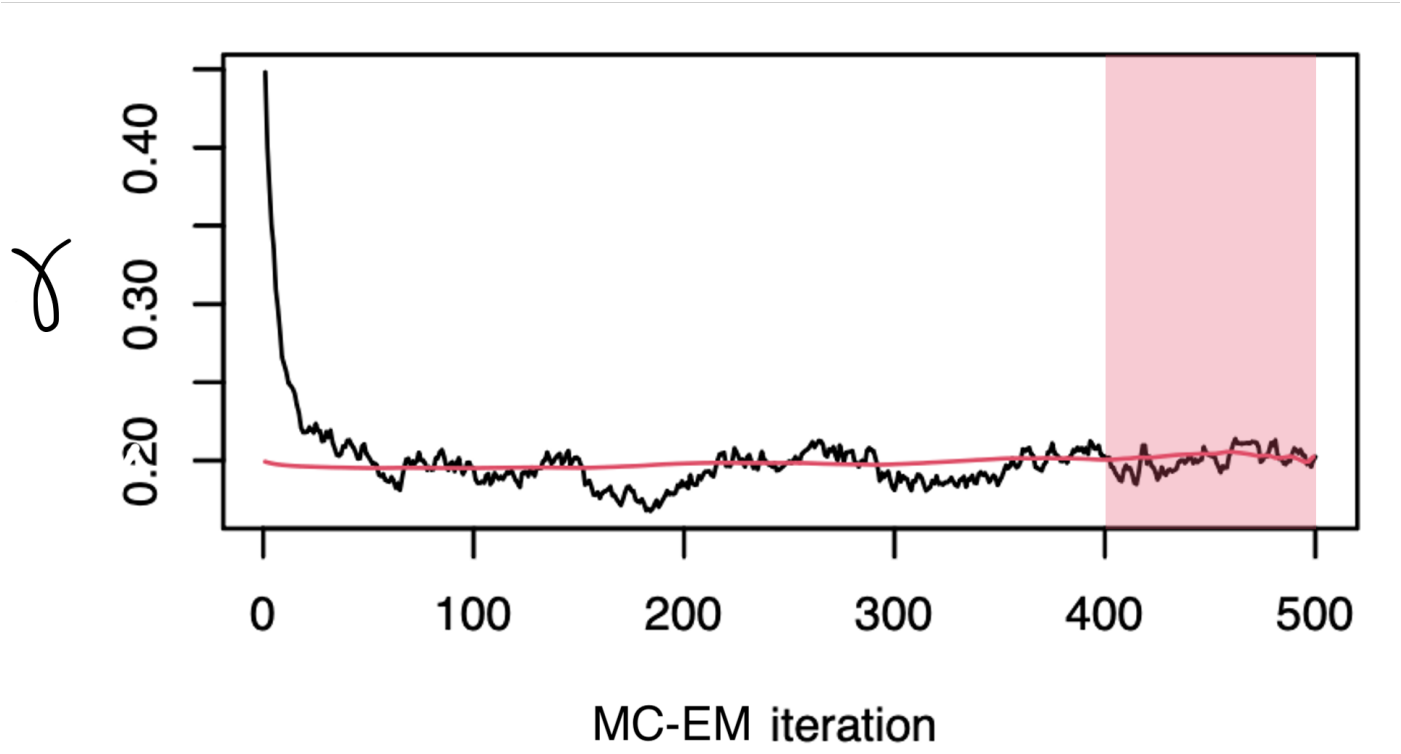


Figure: Convergence of the probability to have a link (truth: 20% of the smallest file) in an MC-EM algorithm with 500 iterations. The last 100 iterations are used to build our estimate of Δ.

Simulations

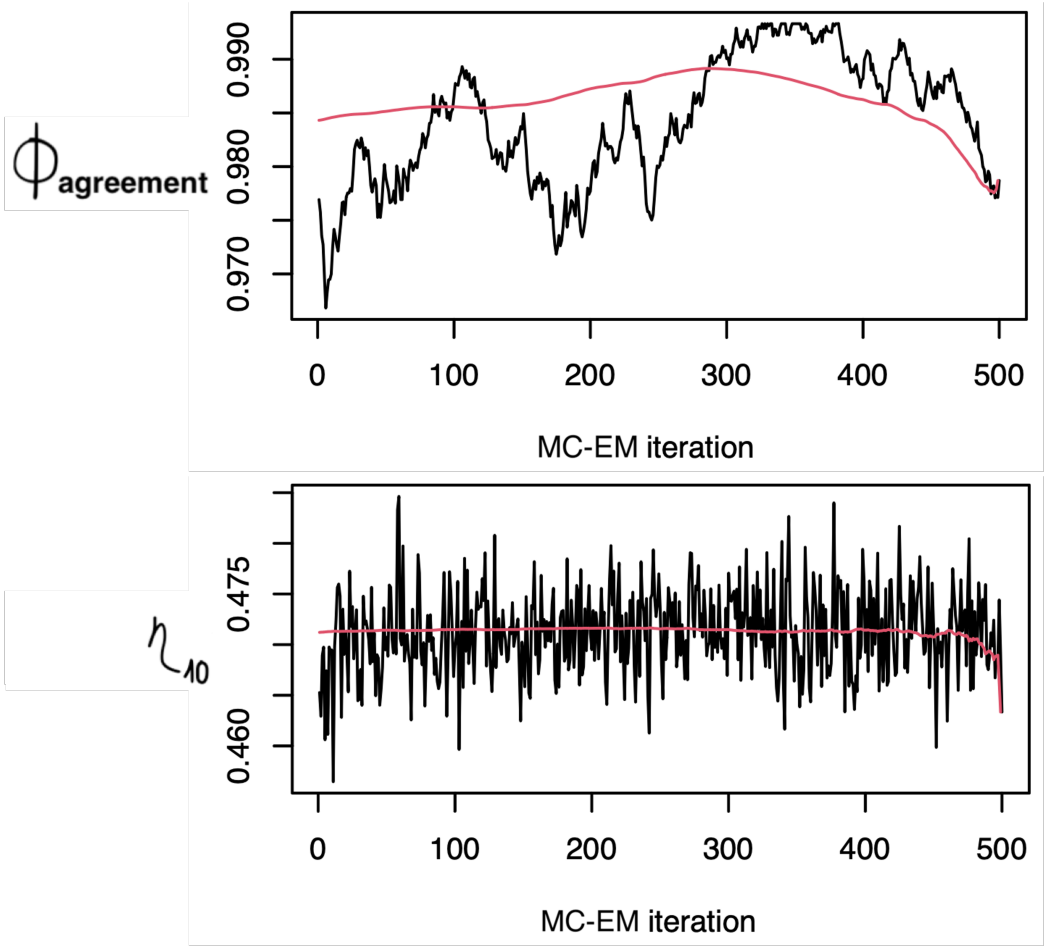


Figure: Convergence of the parameter for agreement among true and registered values (top) and of one parameter for the distribution of true values (bottom).