Detecting duplicates in a homicide registry using a Bayesian partitioning approach

Mauricio Sadinle

Duke University

20/09/2016

Outline

Detecting duplicates in a datafile

- Suppose a datafile is available with a certain number of records.
- If the identifying variable is present in the dataset there is no problem to detect the records that refers to the same entity in the population.
- But...in many cases the identification key is not available.
- Not knowing which are the duplicates can compromise subsequent statistical analyses that make use of that dataset.
- In this work a Bayesian methodology for detecting duplicates is proposed and it is applied to the dataset reporting the homicides during San Salvador civil war (1980-1991).

The standard approaches

- Classical approach doesn't account for the uncertainty of the linkage step and many times it is not transitive.
- Bayesian approach makes the accounting for the uncertainty of the linkage step very natural through the posterior distribution.
- In this work partial agreements between fields' values are taken into account since there exist fields - for instance name or surname - often subjected to typographical errors.

Coreference partition and coreference matrix

- Assume the datafile contains r records and n is the latent number of underlying entities. In other words we can allocate all the r records in n different cells. This allocation is the true latent partition we want to infer on.
- For instance if we have 3 records the true latent partition could be 1,3/2 indicating that records 1 and 3 refer to the same entity while record 2 doesn't have duplicates in the dataset.
- We call coreferents two records referring to the same entity
- We define the coreference matrix (latent) as an $r \times r$ matrix $\mathbf{\Delta}$ such that:

$$\begin{cases} \mathbf{\Delta}_{ij} = 1 \text{ if } (i,j) \text{is a coreferent pair} \\ \mathbf{\Delta}_{ij} = 0 \text{ otherwise} \end{cases}$$

Constrain on the possible coreferent partition

- Detecting the pairs that are obvious nocoreferent reduces tremendously the inferential and computational complexity of the problem.
- Let us define $\mathcal{P}$ the set of pairs for which complete comparisons are computed.
- Within $\mathcal{P}$ many pairs may still be obvious noncoreferent. Then the set of remaining pairs whose coreference status is still unknown is denoted by $\mathcal{C}$ but although the pairs in $\mathcal{P} - \mathcal{C}$ are fixed as noncoreferent their comparation data are used as example of noncoreferent records.
- The possible coreference partition of the file is now constrained to the set $\mathcal{D} = \{\boldsymbol{\Delta} : \Delta_{ij} = 0, \forall (i,j) \notin \mathcal{C}\}$

Other representation of partition and prior distribution

- Representing partitions using matrices is computationally inefficient
- Let us define the r-dimensional vector $\mathbf{Z} = (Z_1, Z_2, ..., Z_r)$ where $Z_i = q$ if record i represents entity q. Then we have: $\mathbf{\Delta_{ij}} = I(Z_i = Z_j)$ where $I()$ is the indicator function.
- Notice that a partition of r elements into n cells has $\frac{r!}{(r-n)!}$ possible labellings.
- It is possible to obtain a flat prior on Δ imposing the following prior on $\mathbf{Z}$:

$$\pi(\mathbf{Z}) \propto \left[\frac{(r - n(\mathbf{Z}))!}{r!}\right] I(\mathbf{Z} \in \mathcal{Z})$$

where $\mathcal{Z} = \{\mathbf{Z} : Z_i \neq Z_j, \forall (i,j) \notin \mathcal{C}\}$

The comparison data

- To compare two records we need to compare the values assumed by the fields of this two records.
- Suppose that the generic field f has $l = 0, 1, ..., L_f + 1$ levels of disagreement. The level 0 is to indicate total agreement.
- Let us define γ_{ij}^{f} the comparison between record i and j concerning the field f
- We say that $\gamma_{ij}^{f} = l$ if the level of disagreement between i and j in the field f is equal to l where $l = 0, 1, 2, ..., L_f + 1$
- Let us define $\boldsymbol{\gamma}_{\mathbf{ij}} = (\gamma_{ij}^{1}, ..., \gamma_{ij}^{f}, ..., \gamma_{ij}^{F})$ where F is the number of fields.

Outline

The model for coreferent and noncoreferent pairs

- It is assumed a different model for coreferent and noncoreferent pairs.
- In particular we can say that:

$$\Gamma_{ij}|\Delta_{ij} = 1 \sim G_1,$$

$$\Gamma_{ij}|\Delta_{ij} = 0 \sim G_0$$

for all $(i, j) \in \mathcal{P}$ where G_1 and G_0 represent the models for coreferent and noncoreferent pairs, respectively.

Joint distribution of the comparison data

- The joint distribution of comparison data can be written as:

$$P(\mathbf{\Gamma} = \gamma | \mathbf{\Delta}, \mathbf{\Phi}) = \prod_{(i,j) \in \mathcal{C}} P_1(\gamma_{\mathbf{ij}} | \Phi_1)^{\mathbf{\Delta_{ij}}} P_0(\gamma_{\mathbf{ij}} | \Phi_0)^{1-\mathbf{\Delta_{ij}}} \times \prod_{(i,j) \in \mathcal{P}-\mathcal{C}} P_0(\gamma_{\mathbf{ij}} | \Phi_0) \quad (1)$$

where $P_1(\gamma_{\mathbf{ij}} | \Phi_1) := P(\mathbf{\Gamma_{ij}} | \Delta_{ij} = 1, \Phi_1)$ and, similarly, $P_0(\gamma_{\mathbf{ij}} | \Phi_0) := P(\mathbf{\Gamma_{ij}} | \Delta_{ij} = 0, \Phi_0)$ with $\mathbf{\Phi} = (\Phi_1, \Phi_0)$ representing a parameter vector of the models G_1 and G_0.

Outline

Missing at random

- It is common to find records with missing fields of information which cause missing comparations for the corresponding record pairs.
- It is assumed that the missing comparation occur at random (MAR). Under this hypothesis it is possible to base the inference on the marginal distribution of the observed comparations and (1) becomes:

$$P(\mathbf{\Gamma}^{obs} = \gamma^{obs} | \mathbf{\Delta}, \mathbf{\Phi}) = \prod_{(i,j)\in\mathcal{C}} P_1(\gamma_{\mathbf{ij}}^{obs} | \Phi_1)^{\mathbf{\Delta}_{\mathbf{ij}}} P_0(\gamma_{\mathbf{ij}}^{obs} | \Phi_0)^{1-\mathbf{\Delta}_{\mathbf{ij}}} \times \prod_{(i,j)\in\mathcal{P}-\mathcal{C}} P_0(\gamma_{\mathbf{ij}}^{obs} | \Phi_0) \quad (2)$$

where $P_1(\gamma_{\mathbf{ij}}^{obs} | \Phi_1) = \sum_{\gamma_{\mathbf{ij}}^{\mathbf{mis}}} P_1(\gamma_{\mathbf{ij}}^{obs}, \gamma_{\mathbf{ij}}^{\mathbf{mis}} | \Phi_1)$

Replacing $\boldsymbol{\Delta}$ with $\mathbf{Z}$ we obtain:

$$P(\boldsymbol{\Gamma}^{obs} = \gamma^{obs} | \mathbf{Z}, \boldsymbol{\Phi}) = \prod_{(i,j) \in \mathcal{C}} P_1(\gamma_{\mathbf{ij}}^{obs} | \Phi_1)^{I(Z_i = Z_j)} P_0(\gamma_{\mathbf{ij}}^{obs} | \Phi_0)^{Z_i \neq Z_j} \times \prod_{(i,j) \in \mathcal{P} - \mathcal{C}} P_0(\gamma_{\mathbf{ij}}^{obs} | \Phi_0) \quad (3)$$

- Let us define $m_{f0} = P_1(\Gamma_{ij}^f = 0)$, $m_{fl} = P_1(\Gamma_{ij}^f = l | \Gamma_{ij}^f > l-1)$ for $0 < l < L_f$ Moreover $u_{f0} = P_1(\Gamma_{ij}^f = 0)$, $u_{fl} = P_1(\Gamma_{ij}^f = l | \Gamma_{ij}^f > l-1)$ for $0 < l < L_f$
- The assumption of the comparison fields being conditionally independent (CI) make easy to explicit $P_1(\gamma_{ij}^{obs}|\Phi_1)$ and $P_0(\gamma_{\mathbf{ij}}^{obs}|\Phi_0)$. In particular:

$$P_1(\gamma_{ij}^{obs}|\Phi_1) = \prod_{f=1}^{F} \left[\prod_{l=0}^{L_f-1} m_{fl}^{I(\gamma_{ij}=l)} (1 - m_{fl})^{I(\gamma_{ij}>l)} \right]^{I_{obs}(\gamma_{ij}^f)} \tag{4}$$

$$P_0(\gamma_{ij}^{obs}|\Phi_0) = \prod_{f=1}^{F} \left[\prod_{l=0}^{L_f-1} u_{fl}^{I(\gamma_{ij}=l)} (1 - u_{fl})^{I(\gamma_{ij}>l)} \right]^{I_{obs}(\gamma_{ij}^f)} \tag{5}$$

Likelihood: Combining (3) with (4) and (5) it is easy to explicit the likelihood for Z and $\Phi = (\mathbf{m}, \mathbf{u})$

Prior on $\mathbf{m}$

$$m_{fl} \sim Uniform(\lambda_{fl}, 1), 0 < \lambda_{fl} < 1$$

$$l = 0, 1, ..., L_f + 1 \text{ and } f = 1, 2, ..., F$$

Prior on $\mathbf{u}$

$$u_{fl} \sim Uniform(0, 1)$$

,

$$l = 0, 1, ..., L_f + 1 \text{ and } f = 1, 2, ..., F$$

The inference is performed via Gibbs Sampling

A simple example

	G.name	F. name	Y	M	D	Mun
R1	JOSE	FLORES	1981	1	29	A
R2	JOSE	FLORES	1981	2	NA	A
R3	JOSE	FLORES	1981	3	20	A
R4	JULIAN ANDRES	RAMOS ROJAS	1986	8	5	B
R5	JILIAM	RMAOS	1986	8	5	B

Table: Y=Year,M=Month,D=Day, Mun=Municipality

Posterior results

- Case 1 (Prior truncation: Given and Family name 0.85, Day and Month 0.85) Posterior concentrated on the partition 1,2,3/4,5
- Case 2 (Prior truncation: Given and Family name 0.85, Day and Month 0.95) Posterior concentrated on the partitions 1,2/3/4,5 and 1/2,3/4,5
- Case 3 (Prior truncation: Given and Family name 0.95, Day and Month 0.85) Posterior concentrated on the partition 1,2,3/4/5
- Case 4 (Prior truncation: Given and Family name 0.95, Day and Month 0.95) Posterior concentrated on the partitions 1,2/3/4/5 and 1/2,3/4/5

THANK YOU!