Some advances on Bayesian record linkage, population size estimation and inference for linked data

Brunero Liseo
(joint research with Andrea Tancredi)

Sapienza Università di Roma

ESSnet ESSnet Data Integration Workshop
INE - Madrid 24-25 November 2011

- Linking data from different sources may be a necessary step in several statistical applications
  - Population size estimation via capture recapture models: official statistics, epidemiology, ecology...
  - Shape analysis: bioinformatics
  - Regression modelling with Y and X's observed in different files sharing some common units
- Linking data aims at producing a larger data set
- Problems arise when datasets do not share a common identifier: units are not labelled.

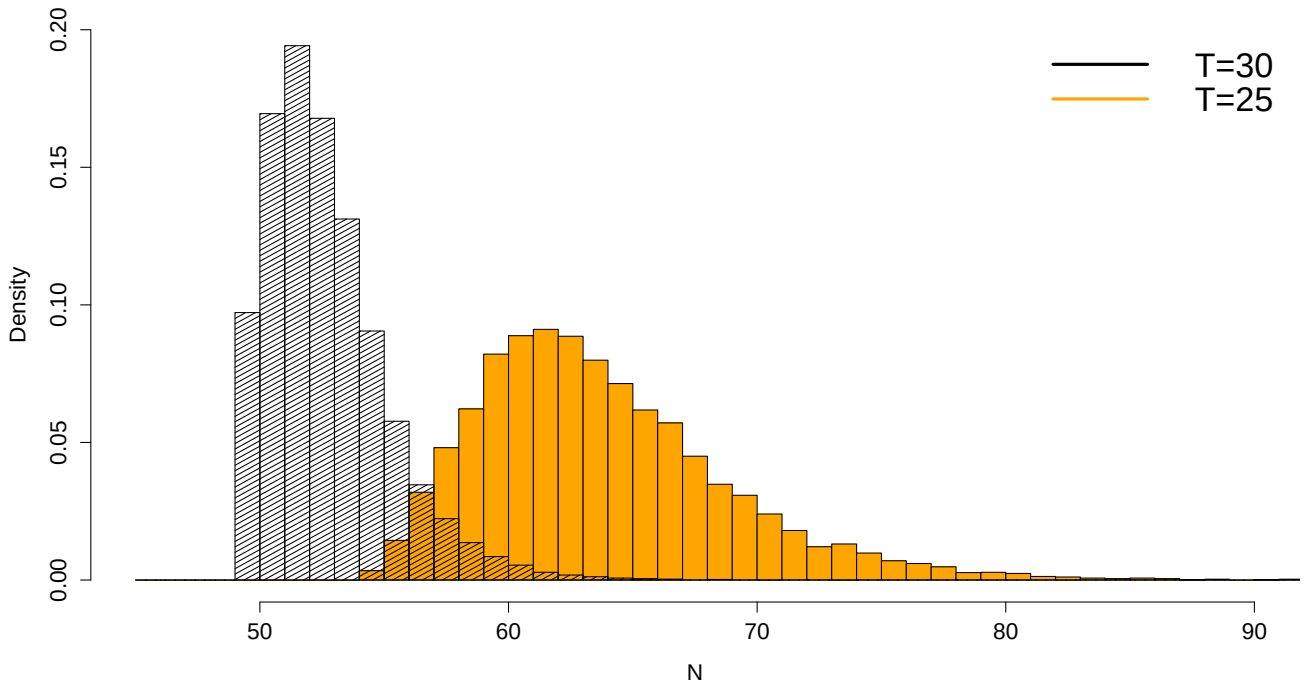
Simple example

- Data set A ($n^A = 34$) contains all the foreigner residents observed in a small census block during the last Italian census population survey
- Data set B ($n^B = 45$) contains all the foreigner residents observed in the same census block during the post enumeration survey - POS.
- Matching variables: first two letters of the *coded surname*, *gender* and *education*.

  The parameter of interest is N, the number of foreigner residents in the census block.

Posterior distribution of N

Our goal is to estimate N by taking into account the *recaptures* uncertainty. Let T be the number of matches. Similar choices of T may produce dramatic different posteriors.



Classic approach to record linkage

Suppose we are given two record configurations x^A and x^B of different sizes n^A and n^B with

$$x^A = (x_1^A, \ldots, x_a^A, \ldots, x_{n^A}^A)' \quad \text{and} \quad x^B = (x_1^B, \ldots, x_b^B, \ldots, x_{n^B}^B)'.$$

Here $x_a^A = (x_a^{A_1}, \ldots, x_a^{A_h})$ and $x_b^B = (x_b^{B_1}, \ldots, x_b^{B_h})$ are the observed values of a categorical random vector $x = (x^1, \ldots, x^h)$ whose support is the set

$$V = \left\{ v_{j_1 j_2, \ldots, j_h} = (v_{j_1}^1, v_{j_2}^2, \ldots, v_{j_h}^h) \quad j_1 = 1 \ldots, k_1; \ldots; j_h = 1, \ldots k_h \right\}.$$

Let M be the set that represents identical units and let U be the set representing different units

$$M = \{(a, b) \in A \times B : a = b\} \quad \text{and} \quad U = \{(a, b) \in A \times B : a \neq b\}.$$

Record linkage analysis is usually performed via the construction of vectors $y_{ab} = (y_{ab}^1, \cdots, y_{ab}^h), a \in A;\ b \in B,$

$$y_{ab}^i = \begin{cases} 1 & x_a^{A_i} = x_b^{B_i} \\ 0 & x_a^{A_i} \neq x_b^{B_i} \end{cases}, \qquad i = 1, \ldots, h.$$

Comparison vectors y'_{ab}s are assumed to be *i.i.d.* with distribution given by the mixture

$$p(y_{ab}|m, u, w) = w \prod_{i=1}^{h} m_i^{y_{ab}^i}(1 - m_i)^{1-y_{ab}^i} + (1 - w) \prod_{i=1}^{h} u_i^{y_{ab}^i}(1 - u_i)^{1-y_{ab}^i}.$$

To perform the record linkage consider

$$\lambda = \frac{P(y_{ab}|(a, b) \in M)}{P(y_{ab}|(a, b) \in U)} = \frac{\prod_{i=1}^{h} m_i^{y_{ab}^i}(1 - m_i)^{1-y_{ab}^i}}{\prod_{i=1}^{h} u_i^{y_{ab}^i}(1 - u_i)^{1-y_{ab}^i}}$$

or the posterior probability $p((a, b) \in M|y_{ab})$

Several extensions of this basic setup have been proposed.
The general approach can be criticized on different grounds.

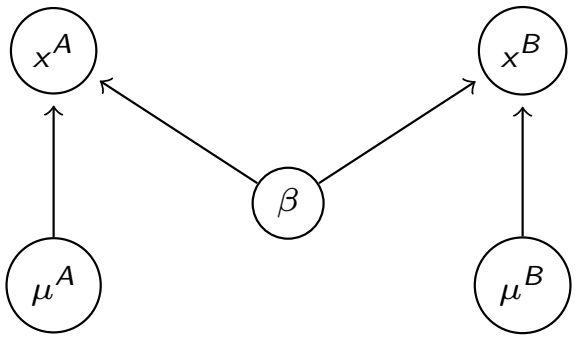
- Decision rules for classifying records as matches are problematic
- Multiple matches are not ruled out
- Sampling information about the X values is not considered
- Comparison vectors are not independent
- The components $y_{ab}^{i}\ i = 1, \ldots, h$ may be not conditionally independent
- Model for categorical variables

A new model for categorical variables

Data sets x^A and x^B contain measurements subject to recording error of a multivariate variable $\mu = (\mu^1, \ldots, \mu^h)$ whose support is V.

Let $\mu^A = (\mu_1^A, \cdots, \mu_a^A, \cdots, \mu_{n^A}^A)'$ and $\mu^B = (\mu_1^B, \cdots, \mu_b^B, \cdots, \mu_{n^B}^B)'$ be the two unobserved matrices with μ_s^S being the unobserved true vector for unit s in sample S, $S = A, B$

Conditionally on their respective true values and a parameter vector $\beta = (\beta_1, \ldots, \beta_h)$, x^A and x^B are mutually independent random vectors

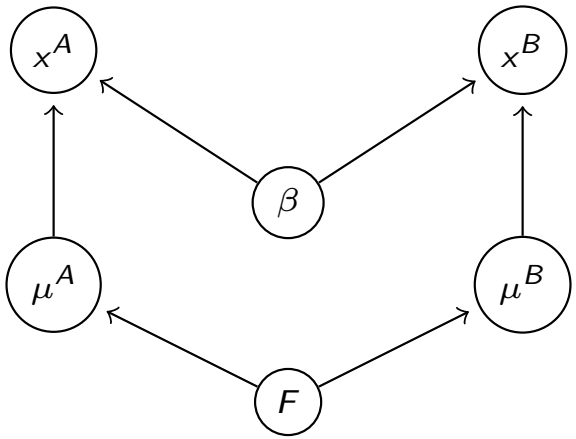


A new model for categorical variables

Conditionally on its true value, each field is a mixture of two components: the first one is concentrated on the true value, and the other one has a uniform spread over $v^i = \{v^i_1, \ldots, v^i_{h_i}\}$

$$p(x^i = v^i_{j_i} | \mu^i = v^i_{j'_i}) = \beta_i \, I_{\{v^i_{j_i} = v^i_{j'_i}\}} + (1 - \beta_i)\frac{1}{k_i} \quad i = 1, \ldots, h$$

μ^A and μ^B are two independent simple random sample drawn without replacement - SRSWR - from a finite population of unknown size N;
To model μ we need to introduce the vector $F = (F_1, \ldots, F_j, \ldots, F_k)$ i.e. the population counts for each element v_j of the set V $(\sum_{j=1}^{k} F_j = N)$



The matching matrix

Consider a $n^A \times n^B$ matching matrix C s. t.

$$C_{ab} = \begin{cases} 1 & \text{if } (a,b) \in M \\ 0 & \text{if } (a,b) \in U \end{cases}$$

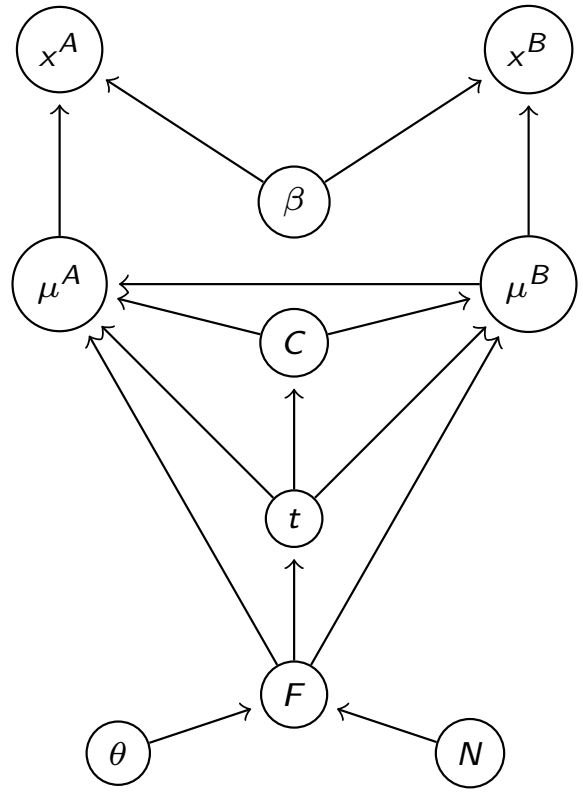
- $\sum_a C_{a\,b} \leq 1 \,\forall\, b = 1, \ldots, n^B$, $\sum_b C_{a\,b} \leq 1 \,\forall\, a = 1, \ldots, n^A$
- $\binom{n^A}{T}\binom{n^B}{T} T!$ different C matrices with $T = \sum_{a\,b} C_{a\,b}$ matches, for all $T \leq \min(n_A, n_B)$.

Further computational semplification comes from the introduction of the vector $t = (t_1, \ldots, t_k)$ denoting, for each element of V, the number of matches taking v_j as the true value. Note $\sum_{j=1}^{k} t_j = T$.

- The conditional distributions $p(\mu^A, \mu^B | C, t, F)$ and $p(t, C|F)$ must reflect the random selection mechanism of the two samples.
  - $p(\mu^A, \mu^B | C, t, F)$: multivariate hypergeometric distribution for the *true* sample counts after eliminating from the population the match counts t_j and fixing the match positions given by C
  - $p(C|t, F)$: uniform distribution on its support, which depends on T; hypergeometric distributions for $t|T, F$ and $T|F$ (it depends on N)
  - Easy to show that "integrating out" t and C one gets two independent *SRSWR*
- Super-population model generating the population counts F
  - Conditionally on the population size N and a probability vector $\theta = (\theta_1, \ldots, \theta_k)$, F is multinomial with size N and parameters θ.
  - the prior for N will be non-informative, i.e. $p(N) \propto 1/N^g$ and θ follows a hyper-Dirichlet distribution

- The conditional distributions $p(\mu^A, \mu^B | C, t, F)$ and $p(t, C|F)$ must reflect the random selection mechanism of the two samples.
  - $p(\mu^A, \mu^B | C, t, F)$: multivariate hypergeometric distribution for the *true* sample counts after eliminating from the population the match counts t_j and fixing the match positions given by C
  - $p(C|t, F)$: uniform distribution on its support, which depends on T; hypergeometric distributions for $t|T, F$ and $T|F$ (it depends on N)
  - Easy to show that "integrating out" t and C one gets two independent *SRSWR*
- Super-population model generating the population counts F
  - Conditionally on the population size N and a probability vector $\theta = (\theta_1, \ldots, \theta_k)$, F is multinomial with size N and parameters θ.
  - the prior for N will be non-informative, i.e. $p(N) \propto 1/N^g$ and θ follows a hyper-Dirichlet distribution

The final DAG structure



Bayesian implementation

A Metropolis within Gibbs algorithm has been used to simulate from the posterior distribution

$$p(\mu^A \mu^B, \beta, t, F, N, \theta | x^A, x^B).$$

Gibbs steps are based on the following block updating

$$\begin{aligned} \mu^A, \mu^B, t &\mid F, N, \theta, \beta \\ F, N &\mid \mu^A, \mu^B, t, \theta, \beta \\ \theta &\mid \mu^A, \mu^B, t, F, N, \beta \\ \beta &\mid \mu^A, \mu^B, t, F, N, \theta. \end{aligned}$$

Notice that C is missing in the above scheme.
When the matching matrix C is itself a quantity of interest, one should add a draw from the conditional distribution at each iteration of the algorithm

$$p(C | \mu^A \mu^B, \beta, t, F, N, \theta, x^A, x^B).$$

This entails a sort of simple Monte Carlo estimate for the posterior of C

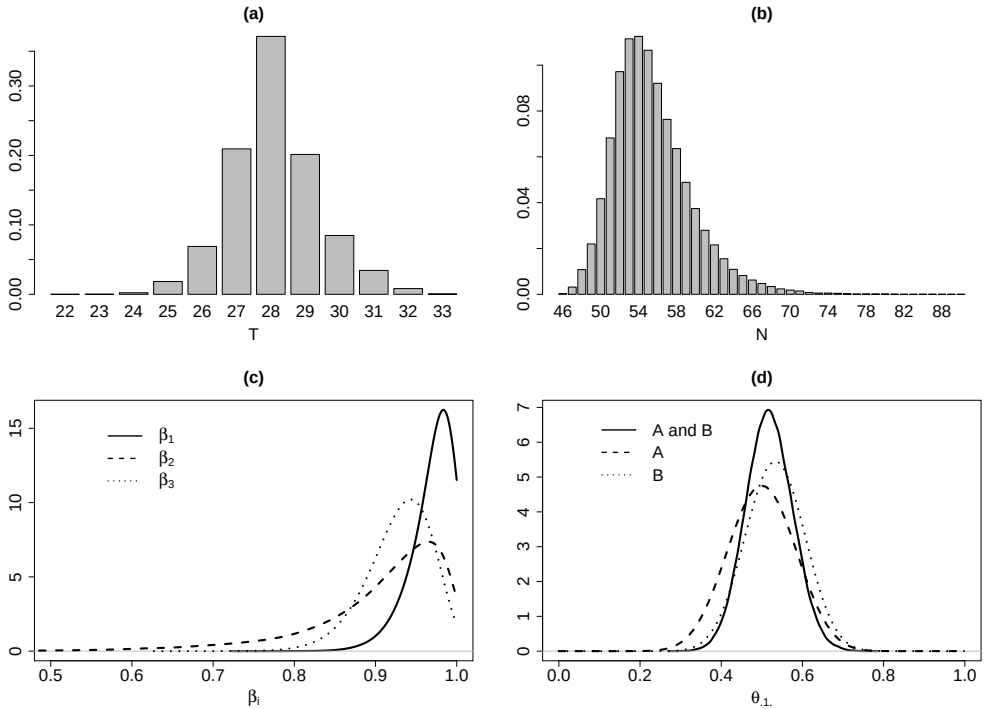
Application

Back to the motivating example:

For the coded surname we have 339 categories while for the level of education we have 17 categories. The total number of entries in the set V is $k = 339 \times 2 \times 17 = 11526$.

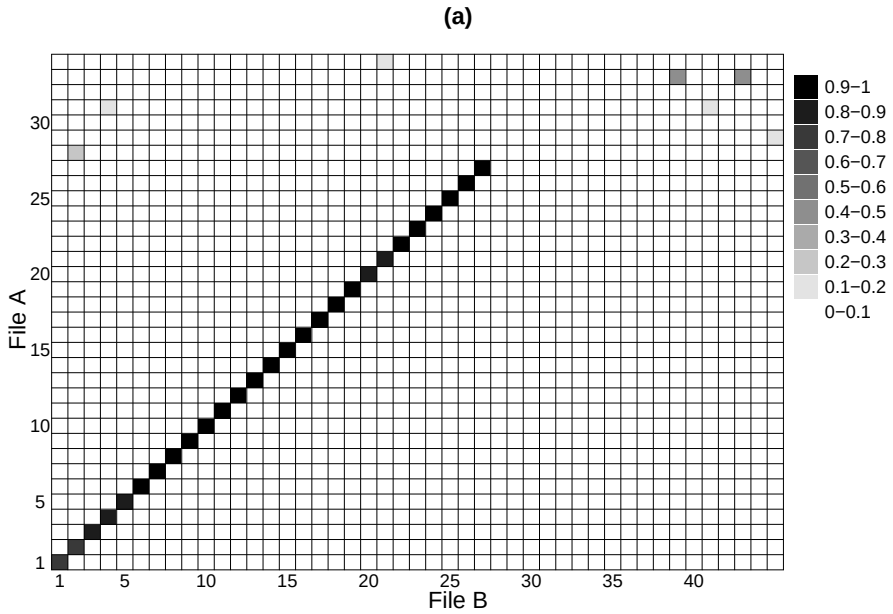
The hyperparmeter g for the prior $p(N)$ has been set equal to 2. The hyper-Dirichlet structure of the prior for θ allows us to assume specific dependence structure among the covariates. Here we assume that the coded surname is independent of sex and education level.

Some results



Posterior distributions for the match number T, the population size N, the parameters β_i $(i = 1, \ldots, 3)$ and $\theta_{\cdot 1 \cdot}$, that is the probability to select a male in the census block for the super-population model.

Posterior Probabilities of matches



$p(C_{ab} = 1|x^A, x^B)$

Alternative approach 1: the Jaro constrained model

- Slight modification of the Bayesian approach proposed by Larsen[2005], where y_{ab} is marginally distributed as a multivariate Bernoulli mixture model and the matching matrix C satisfies the constraints $\sum_a C_{ab} \leq 1$ and $\sum_b C_{ab} \leq 1$.
- Uniform priors for m and u. For the matching matrix C, same prior distribution as before.
- The posterior distribution for the parameters (m, u, C, N) can be easily simulated by using Gibbs steps for $[m|u, C, N]$, $[u|m, C, N]$ and $[N|u, m, C]$. To update the matching matrix C, one can use a single Metropolis-Hastings step

Alternative approach 2: the hybrid approach

- Standard analysis of the Jaro model (Bayesian or classical)
- Plug-in the model estimates into $p((a,b) \in M | y_{ab})$
- Maximization of the function

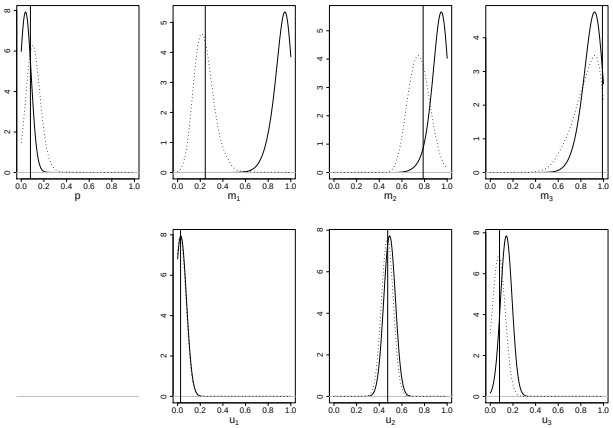
$$\sum_{a=1}^{n^A} \sum_{b=1}^{n^B} z_{ab} \log \frac{\prod_{i=1}^{k} (\hat{m}^{y_{ab}^i} (1-\hat{m})^{1-y_{ab}^i})}{\prod_{h=1}^{k} (\hat{u}^{y_{ab}^i} (1-\hat{u})^{1-y_{ab}^i})}$$

  subject to the constraints $\sum_{a=1}^{\nu_A} z_{ab} \leq 1 \ \forall b$, $\sum_{b=1}^{\nu_B} z_{ab} \leq 1 \ \forall a$ and $z_{ab} \in \{0,1\} \ \forall (a,b)$ in order to avoid multiple matches.

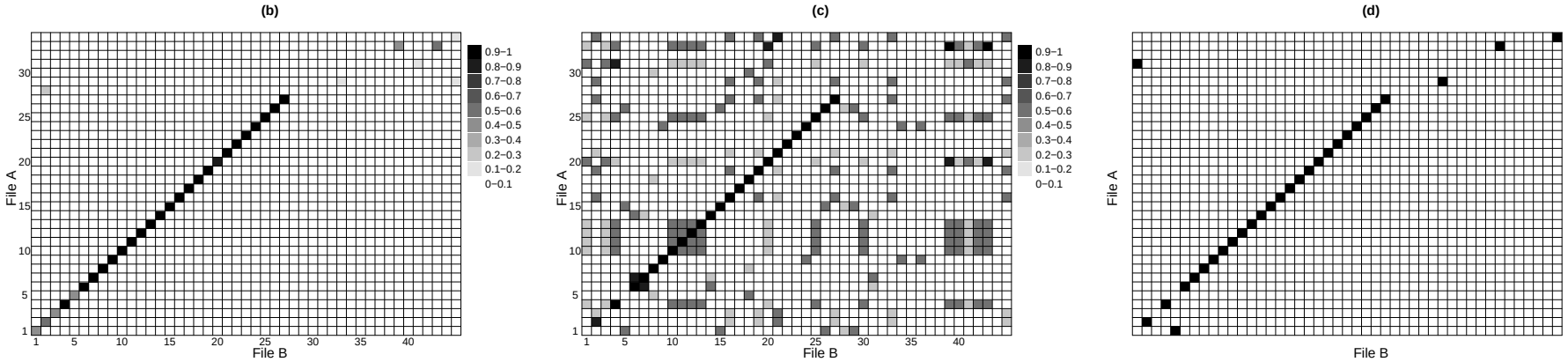
- Plug-in the estimated number of matches into a capture-recapture model (for example the hypergeometric model) for estimating N

y_{ab}	frequency	$p((ab) \in M \mid y_{ab})$	λ
$(0,0,0)$	659	0.00	0.01
$(1,0,0)$	20	0.01	0.14
$(0,1,0)$	601	0.00	0.04
$(1,1,0)$	13	0.05	0.58
$(0,0,1)$	78	0.23	3.43
$(1,0,1)$	8	0.80	45.20
$(0,1,1)$	126	0.56	14.81
$(1,1,1)$	25	0.94	194.97

Data summaries for the alternative approaches

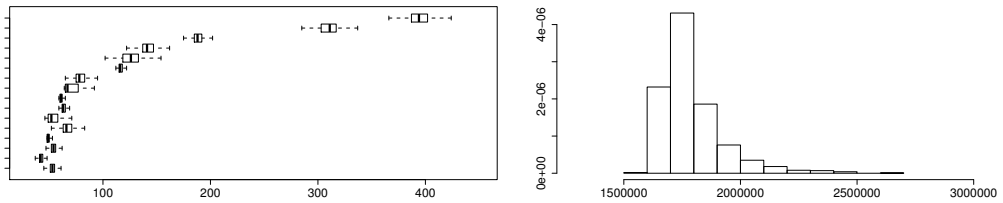


Posterior distributions for the parameters of the mixture model

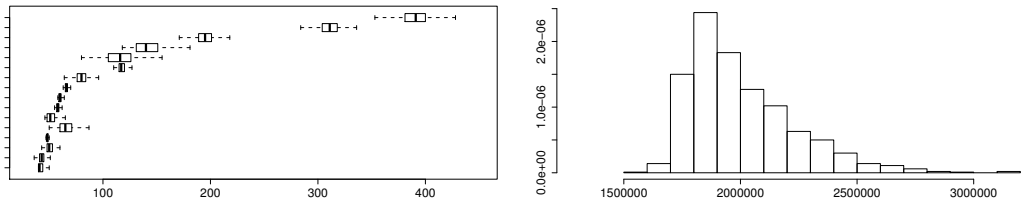


Posterior probabilities of being a match and estimated matching matrix

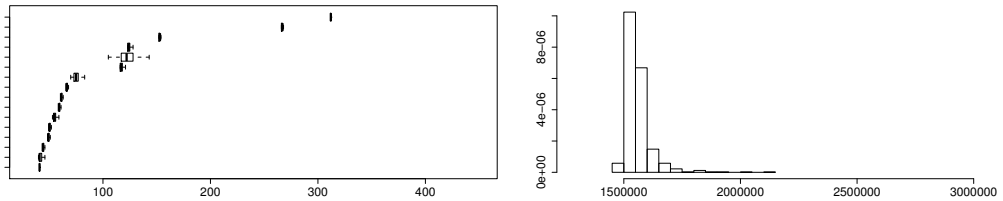
Comparison on all the Italian blocks (4 key variables)



new model



Jaro constrained



hybrid approach

Posterior distribution for the block sizes N_b (left panels) and for $\sum_b N_b$ (right panels)

Matching multivariate normal data

Liseo and Tancredi (2011, JOS)

- Suppose we have $x_a^A \sim N_h(\mu_a^A, \Gamma)$ and $x_b^B \sim N_h(\mu_b^B, \Gamma)$ independently for $a = 1, \ldots, n^A$ and $b = 1, \ldots, n^B$.
- μ_a^A and μ_b^B are two vectors containing the *true values* of the observed variables for individuals $a \in A$ and $b \in B$.
- Given $C_{ab} = 0$, μ_a^A and μ_b^B are two independent $N_h(\theta, \Sigma)$ variates
- Given $C_{ab} = 1$, μ_a^A and μ_b^B take the same value: $\mu_{ab}^{AB} \sim N_{2h}(\theta_2, \Sigma_2)$ where

$$\mu_{ab}^{AB} = \begin{bmatrix} \mu_a^A \\ \mu_b^B \end{bmatrix} \theta_2 = \begin{bmatrix} \theta \\ \theta \end{bmatrix} \quad \text{and} \quad \Sigma_2 = \begin{bmatrix} \Sigma & \Sigma \\ \Sigma & \Sigma \end{bmatrix}$$

- The joint conditional $p(\mu^A, \mu^B | C, \theta, \Sigma)$ is

$$\prod_{a:\, C_{a.}=0} \phi_h(\mu_a^A | \theta, \Sigma) \times \prod_{b:\, C_{.b}=0} \phi_h(\mu_b^B | \theta, \Sigma) \times \prod_{a\, b:\, C_{ab}=1} \phi_{2h}(\mu_{ab}^{AB} | \theta_2, \Sigma_2)$$

- The matching matrix C given T, n^A, n^B is uniformly distributed on its support
- Model M_t for n^A, n^B and the unobserved recaptures $T = \sum_{ab} C_{ab}$

$$n^B | n^A, T, N, p^A, p^B \sim T + \text{Binomial}(N - n^A, p^B)$$

$$n^A | T, N, p^A, p^B \sim T + \text{Binomial}\left(N - T, \frac{p^A(1 - p^B)}{(1 - p^A\, p^B)}\right)$$

$$T | N, p^A, p^B \sim \text{Binomial}(N, p^A\, p^B)$$

BAYESIAN INFERENCE

- Prior independence for $N, p, \Sigma, \theta, \Gamma$ with non-informative or *conditionally conjugate* prior distributions: $p(N) \propto 1/N$, $p \sim Beta$, $\theta \sim$ Multivariate Normal, $\Sigma \sim$Wishart and Γ diagonal with its elements Gamma distributed
- Posterior simulation by a Metropolis whitin Gibbs algorithm after integrating out analytically μ: $p(x^A, x^B | C, \theta, \Sigma, \Gamma)$ is equal to

$$\prod_{a:\, C_{a.}=0} \phi_h(x_a^A | \theta, \Sigma + \Gamma) \times \prod_{b:\, C_{.b}=0} \phi_h(x_b^B | \theta, \Sigma + \Gamma) \times \prod_{a\, b:\, C_{ab}=1} \phi_{2h}(x_{ab}^{AB} | \theta_2, \Psi)$$

where $\Psi = \begin{bmatrix} \Sigma + \Gamma & \Sigma \\ \Sigma & \Sigma + \Gamma \end{bmatrix}$

- Updating of the matching matrix with three moves: add, delete, or switch a match. Main references: Lindley (1977) BKA and Green and Mardia (2006) BKA

Inference with linked data

Hierarchical Bayesian models can be used for the general problem of inference with linked data

- Suppose dataset A contains x^A's and an extra variable Y. Dataset B contains x^B's and some other variables Z's.
- Goal: fit a statistical model linking Y to Z, say $y|z$.
- For simplicity assume $(y, z) \perp\!\!\!\perp \mathbf{x} \mid C$.

$$
\begin{aligned}
p(y, z|C) &= \prod_{a:C_{a.}=0} p(y_a) \prod_{b:C_{.b}=0} p(z_b) \prod_{a\,b:C_{a\,b}=1} p(y_a, z_b) \\
&\propto \prod_{a:C_{a.}=0} p(y_a) \prod_{a\,b:C_{a\,b}=1} p(y_a|z_b)
\end{aligned}
$$

Linear Regression

Scheuren and Winkler (1997) and Lahiri and Larsen (2005) have considered the linear regression problem

- Bayesian analysis allows for several different approaches for this problem.

1. Plug in a point estimate $\hat{C}$ of C and use it in the following regression analysis
2. Include the regression step into the MCMC algorithm. This can be done by noticing that

$$C'Y = C'CX\beta + C'\varepsilon.$$

   (not easy to implement)
3. Hibrid approach; at each iteration of the algorithm, after drawing C, generate a value for (β, σ).

Linear Regression

Scheuren and Winkler (1997) and Lahiri and Larsen (2005) have considered the linear regression problem

- Bayesian analysis allows for several different approaches for this problem.

1. Plug in a point estimate $\hat{C}$ of C and use it in the following regression analysis
2. Include the regression step into the MCMC algorithm. This can be done by noticing that

$$C'Y = C'CX\beta + C'\varepsilon.$$

   (not easy to implement)

3. Hibrid approach; at each iteration of the algorithm, after drawing C, generate a value for (β, σ).

Linear Regression

Scheuren and Winkler (1997) and Lahiri and Larsen (2005) have considered the linear regression problem

- Bayesian analysis allows for several different approaches for this problem.

1. Plug in a point estimate $\hat{C}$ of C and use it in the following regression analysis
2. Include the regression step into the MCMC algorithm. This can be done by noticing that

$$C'Y = C'CX\beta + C'\varepsilon.$$

   (not easy to implement)
3. Hibrid approach; at each iteration of the algorithm, after drawing C, generate a value for (β, σ).

Small scale Simulation

Population size: N=100, Sample sizes: $n_a = 80, n_b = 80$. standard 3 independent key variables generated according to the previous model (measurement error $\beta = .1$)
we add to file B a random variable $w_b \sim N(0,1)$ for $b = 1, \ldots, 80$.
In file A, for each pair (a, b) with $\sum_b C_{ab} = 1$ we generate

$$z_a \sim N(\beta_0 + \beta_1 w_b, \sigma)$$

with $\beta_0 = \beta_1 = \sigma = 1$.
For each unit a such that $\sum_b C_a b = 0$ we nave generated

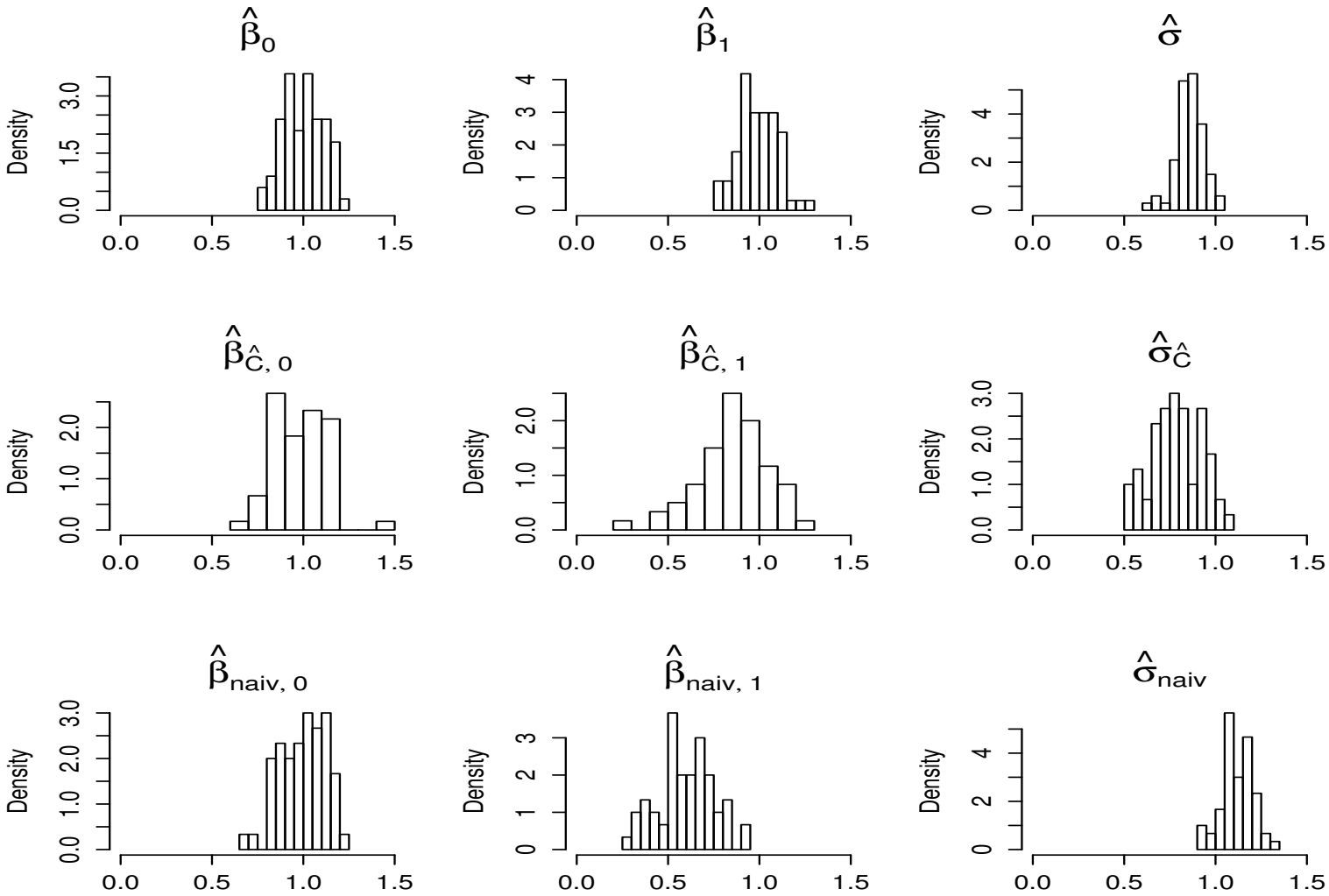
$$z_a \sim N(\beta_0 + \beta_1 \tilde{w}_a, \sigma)$$

with $\tilde{w}_a \sim N(0,1)$

200 replications

We have compared three different strategies, namely

1. sampling distributions of MLE for the three parameters $(\beta_0, \beta_1, \sigma)$ using the true matches (benchmark) (1st line)
2. sampling distributions of MLE for the three parameters $(\beta_0, \beta_1, \sigma)$ using the point estimate of C. (Two-steps Bayesian standard solution) (2nd line)
3. histograms of the means of the MLE estimates evaluated at each iteration of the MCMC algorithm (Naive Bayesian solution) (3rd line)


$\hat{\beta}_0$
$\hat{\beta}_1$
$\hat{\sigma}$
$\hat{\beta}_{\hat{C},\,0}$
$\hat{\beta}_{\hat{C},\,1}$
$\hat{\sigma}_{\hat{C}}$
$\hat{\beta}_{naiv,\,0}$
$\hat{\beta}_{naiv,\,1}$
$\hat{\sigma}_{naiv}$
Density
0.0
0.5
1.0
1.5

Conclusions

Discussion: Bayes and statistics from public domain

- Pragmatic Bayesian approach with several subjective elements
- Objective priors may be used, but subjective elements on hierarchical structures and likelihood functions still remain
- Rebut the idea that frequentist are objective and more appropriate in public domain [Fienberg 2011 Statistical Science]

Possible extensions

- More adequate capture-recapture models should be used (eterogeneity/clustering). Multiple captures
- Remove the assumption of conditional independence of the errors given the true values
- Different models for the misspecified record fields

Some references

- Fienberg. (2011) Bayesian models and methods in public policy and government settings. *Statistical Science*
- Green, Mardia. (2006). Bayesian alignmennt using hierarchical models, with application to protein bioinformatics. *Biometrika*
- Lahiri, Larsen. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*
- Larsen. (2005). Advances in record linkage theory: hierarchical Bayesian record linkage. *ASA proceedings*
- Lindley. (1977). A problem in forensic science. *Biometrika*
- Liseo, Tancredi. (2011). Bayesian estimation of population size via linkage of multivariate normal data sets. *Journal of Official Statistics*, 2011, Vol. 27 No. 3, pp. 491–505
- Scheuren, Winkler (1997). Regression analysis of data files that are computer matched, II, Survey Methodology, 23, 157-165.
- Tancredi, Liseo. (2011). A hierarchical Bayesian approach to record linkage and population size estimation. *Annals of Applied Statistics*, Vol. 5, No. 2B, 1553–1585