

# Hierarchical Bayesian Record Linkage and Regression in Linked Files

Michael D. Larsen

Department of Statistics,  
Now: The University of Chicago,  
August: Iowa State University  
*larsen@galton.uchicago.edu*

Machine Learning, Statistics, and Discovery  
8:00 pm, June 23, 2003  
Snowbird Ski and Summer Resort, Utah

# Outline

1. Record Linkage Problem
2. Data
3. Mixture Models for Classifying Record Pairs
4. A Bayesian approach
5. Regression of linked data
6. Conclusions and Future work

## Record Linkage/File matching problem

	File A	File B
	matching variables <hr/> $v_1 \dots v_K$ <hr/> $X$	matching variables <hr/> $Y$ <hr/> $w_1 \dots w_K$
record $a$	<hr/> <hr/> <hr/> <hr/>	<hr/> <hr/> <hr/> <hr/>
		record $b$

For each pair of records, comparison vector:

$$\gamma(a, b)' = \{\gamma_k(a, b), k = 1, \dots, K\}$$

where  $\gamma_k(a, b) = 1$  if  $v_k(a) = w_k(b)$  and 0 otherwise.

$$A \times B = M \cup U \text{ (Fellegi, Sunter 1969)}$$

$M$ : true matches       $U$ : true nonmatches (unmatched)

Comparison Vectors					Number of Pairs		
$\gamma$					$M$	$U$	Total
1	1	1	1	1	511	4	515
1	1	1	1	0	268	19	287
1	1	1	0	1	115	10	125
1	1	1	0	0	58	54	112
1	1	0	1	1	49	10	59
⋮							
⋮							
0	0	0	0	1	3	1150	1153
0	0	0	0	0	1	9579	9580

Unknown Probabilities:

$P(M)$  and  $P(U)$ ,  $P(\gamma|M)$  and  $P(\gamma|U)$

$\Lambda = P(\gamma|M)/P(\gamma|U)$ , a likelihood ratio

The Fellegi-Sunter (1969) procedure is as follows:

Declare  $(a, b)$  to be a match if  $\Lambda >$  upper cutoff.

Declare  $(a, b)$  to be a nonmatch if  $\Lambda <$  lower cutoff.

Otherwise send  $(a, b)$  to clerical review.

The cutoffs are determined to minimize the amount of clerical review at pre-set error rates.

# The Use of Mixture Models for Record Linkage

Larsen, Rubin JASA March 2001

- Unclassified comparisons arise from a mixture:

$$P(\gamma) = \sum_{g=1}^G \pi_g P(\gamma | \text{class } g).$$

- Maximum Likelihood Estimation via EM/ECM.
- The conditional independence or latent class model:

$$P(\gamma | \text{class } g) = \prod_{k=1}^K P(\gamma_k | \text{class } g)$$

- The data:

counts in  $2^K$  table.

## Mixture Models, continued

Let  $z_{ig} = 1$  if pair  $i$  is from mixture class  $g$  and 0 otherwise.

$$P(z_{ig} = 1 | \gamma, \text{parameters}) = \frac{\pi_g P(\gamma | \text{class } g)}{\sum_{h=1}^G \pi_h P(\gamma | \text{class } h)}$$

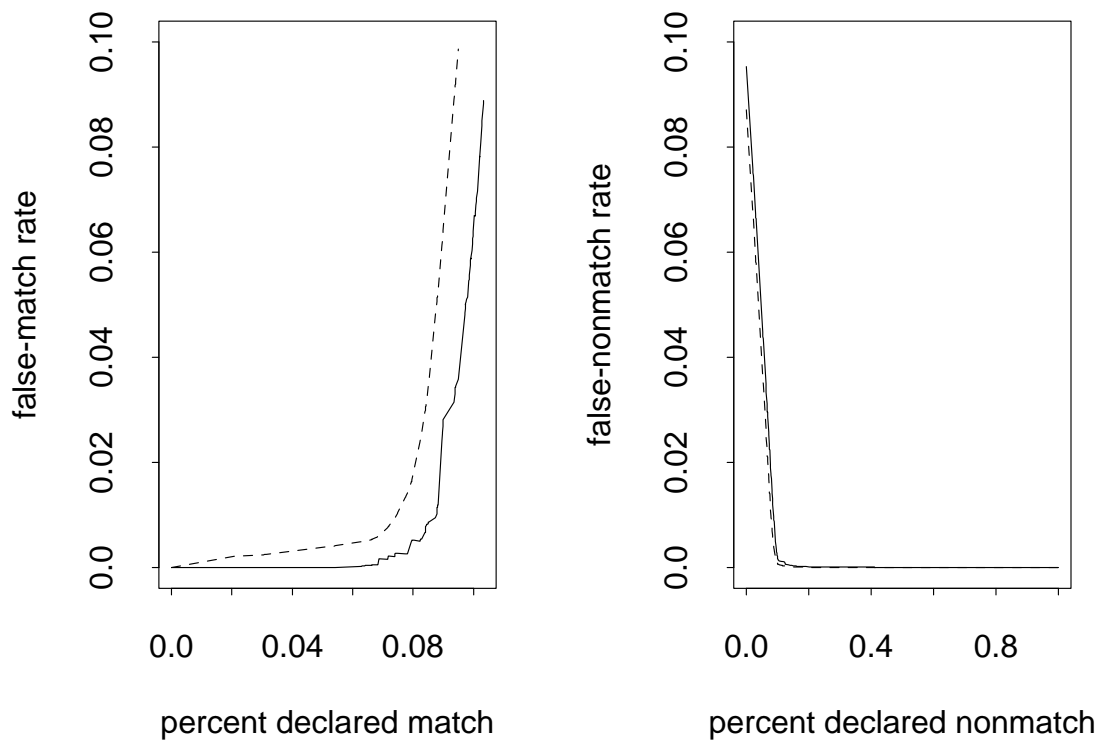
Fellegi-Sunter Ratio:

$$p(\gamma | M) / p(\gamma | U) = p(M | \gamma) p(M) / p(U | \gamma) p(U)$$

Estimated Error Rates:

$$\sum_{i=1}^{i'-1} p(\gamma_i | U) < \mu \leq \sum_{i=1}^{i'} p(\gamma_i | U)$$
$$\sum_{j=j'}^n p(\gamma_j | M) \geq \lambda > \sum_{j=j'+1}^n p(\gamma_j | M),$$

Figure 1: *False-match and false-nonmatch rates from fitting a three-class conditional-independence mixture to D88a.*



## A Bayesian Approach

- Bayesian mixture model approach for record linkage.

$$(P_\gamma, \gamma \in \Gamma | M) \sim \text{Dirichlet}(\alpha_\gamma^M)$$

$$(P_\gamma, \gamma \in \Gamma | U) \sim \text{Dirichlet}(\alpha_\gamma^U)$$

$$p_M \sim \text{Beta}(\alpha, \beta)$$

- In the latent class approach:

$$P(\gamma_k = 1 | M) \sim \text{Beta}(\alpha_{Mk}, \beta_{Mk})$$

$$P(\gamma_k = 1 | U) \sim \text{Beta}(\alpha_{Uk}, \beta_{Uk})$$

$$k = 1, \dots, K$$

$$p_M \sim \text{Beta}(\alpha, \beta)$$

- **Prior Information**

Parameters of Prior Distribution as *prior counts*

Use combination of *ideal data*

and (scaled) *classified data* from a similar site.

- **Simulation of the Posterior Distribution**

Computation of posterior distributions can be accomplished via Data Augmentation or Gibbs sampling.

Cycle between two steps:

- Sample the vectors of indicators
- Draw the parameters

- **Alternative Bayesian approach**

Another Bayesian approach is considered by Fortini, Liseo, Nuccitelli, and Scanu (2000) **ISBA Proceedings**. In their paper they average over the mixture model parameters.

These authors rely heavily on the prior distribution rather than making modeling assumptions or placing restrictions on the parameters. Future work will compare the two approaches.

## Use of Bayesian Output for Record Linkage

- Posterior intervals for Parameters via simulation.
- Posterior intervals for Percent Matches at stated error levels.
- Posterior envelope about the error rate curves.

→ Conservative procedure:

use upper limit of 95% posterior interval

rather than center of distribution.

## Estimation of Regression Coefficients

Consider the following regression model

$$y_i = x_i' \beta + \epsilon_i, i = 1, \dots, n, \quad (1)$$

- $x_i = (x_{i1}, \dots, x_{ip})'$  is a vector of  $p$  known covariates.
- $E(\epsilon_i) = 0$ ,  $Var(\epsilon_i) = \sigma^2$ , and  $cov(\epsilon_i, \epsilon_j) = 0$  for  $i \neq j$ ,  $i, j = 1, \dots, n$ .
- The response ( $X$ ) is in file  $A$ , the covariates ( $Y$ ) are in file  $B$ , and the two files are linked imperfectly.

**Note:** The true pairs  $(x_i, y_i)$  are not observable.

## A Model for $z_i$ 's:

$$z_i = \begin{cases} y_i & \text{with probability } q_{ii} \\ y_j & \text{with probability } q_{ij} \text{ for } i \neq j, \end{cases} \quad (2)$$

where  $\sum_{j=1}^n q_{ij} = 1$ ,  $i, j = 1, \dots, n (i \neq j)$ .

## A naive estimator of $\beta$ :

$$\hat{\beta}_N = (X'X)^{-1}X'Z,$$

where  $X = (x'_1, \dots, x'_n)'$  and  $Z = (z_1, \dots, z_n)'$ .

is biased.

- An improved estimator was presented by

Scheuren and Winkler (1993) *Survey Methodology*

- An iterative procedure was presented by

Scheuren and Winkler (1997) *Survey Methodology*

- Lahiri and Larsen (2003; submitted *JASA*) developed an alternative estimator of  $\beta$  and its SE.

★ SW (1993) and LL (2003) use estimated probabilities of matching for adjustment.

## Larsen, Lahiri unbiased estimator of $\beta$

Lahiri and Larsen (2003), under revision

Note that under the model described by (1) and (2),

$$E(z_i) = w_i' \beta, \quad i = 1, \dots, n,$$

where  $w_i = \sum_{j=1}^n q_{ij} x_j$ ,  $i = 1, \dots, n$ .

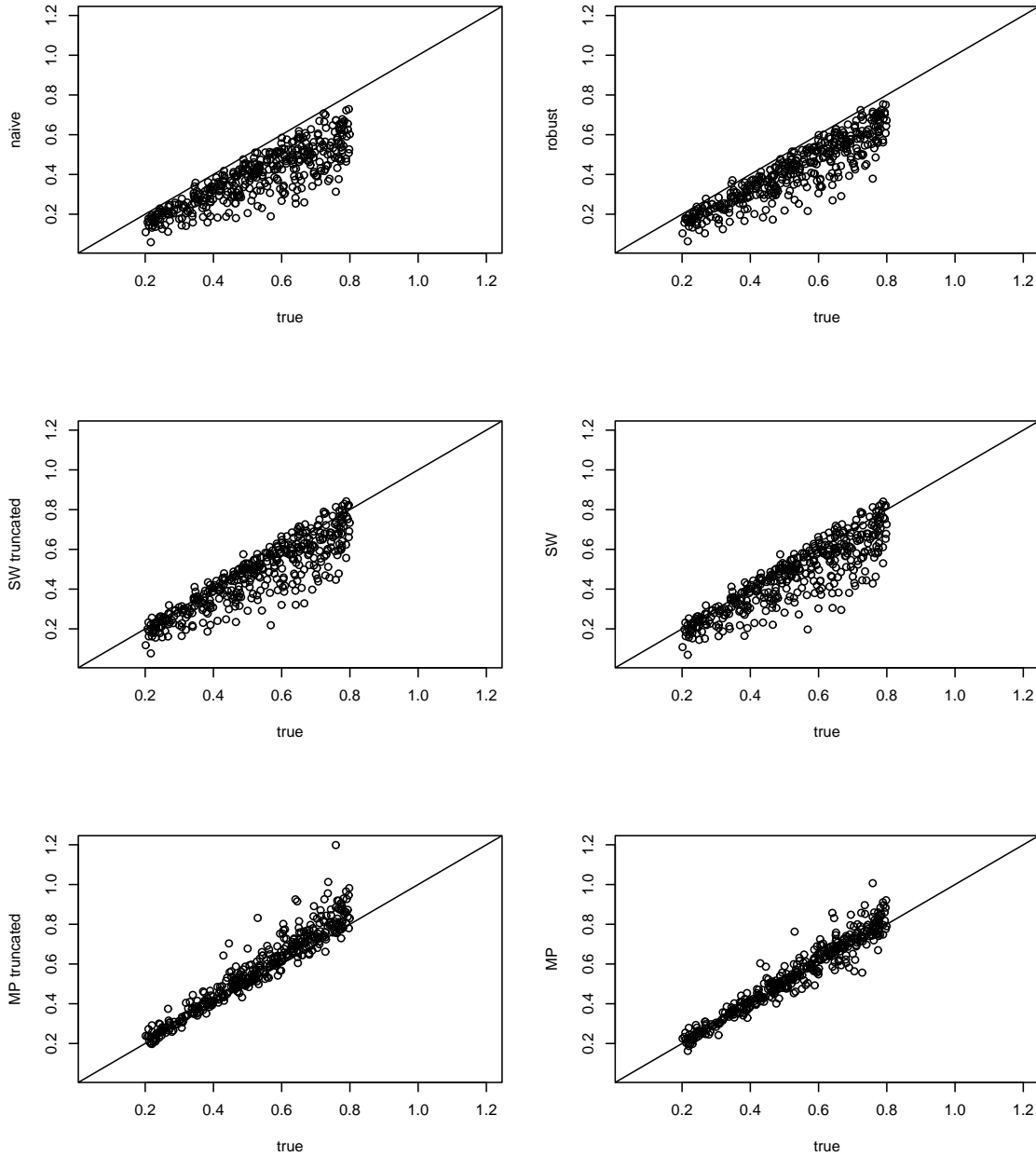
Thus, an unbiased estimator of  $\beta$  is given by

$$\hat{\beta}_U = (W'W)^{-1}W'Z,$$

where  $W' = (w_1, \dots, w_n)$ .

- We also have an estimator of the variance of  $\hat{\beta}_U$ .

Figure 2: *Comparison of Six Estimators on Four Hundred Data Sets, second set of simulation conditions. Plots of Naive, Robust, Scheuren-Winkler (truncated and full), Lahiri-Larsen (truncated and full) estimators versus the truth. Diagonal lines have slope 1.*



## **(Partially) Bayesian Regression of Linked Data**

- Simulate the record linkage parameters from their posterior distribution.
- Compute the regression estimator for each draw from the record linkage parameter distribution.
- Report the median (and 95% central interval) of regression estimates.
- **Advantage:** Better coverage of simulation slopes.

◇ Future work:

Bayesian regression version of Larsen-Lahiri (2003).

That is, specify a prior distribution on the regression model parameters.

## Blocking and 1-1 assignment

- Blocking

Not all record pairs  $(a, b)$  are compared to one another; record pairs are compared only within blocks of records.

Characteristics vary across blocks.

If sample sizes within blocks were not small, one could imagine applying the Fellegi-Sunter procedure separately in each block.

Let  $b = 1, \dots, B$  index blocks

- 1-1 assignment.

If no duplicates ...

In block  $b$ , let  $I_{ijb} = 1$  if subjects  $i$  in file A and  $j$  in file B are links, 0 otherwise.

$$\sum_{i \in \text{block}b} I_{ijb} \leq 1$$

$$\sum_{j \in \text{block}b} I_{ijb} \leq 1$$

$$\sum_{i \in \text{block}b} I_{ijb} = \sum_{j \in \text{block}b} I_{ijb}$$

- Two Logical constraints:

$$\begin{aligned} \# \text{ linked pairs in a block} &\leq \\ &\min(\#A, \#B \text{ in the block}) \end{aligned}$$

$$P(\gamma_k|M) \geq P(\gamma_k|U), k = 1, \dots, K$$

## Impact of Inequality Constraints:

**Decrease the number of estimated matches**

**Make probability estimates more extreme**

---

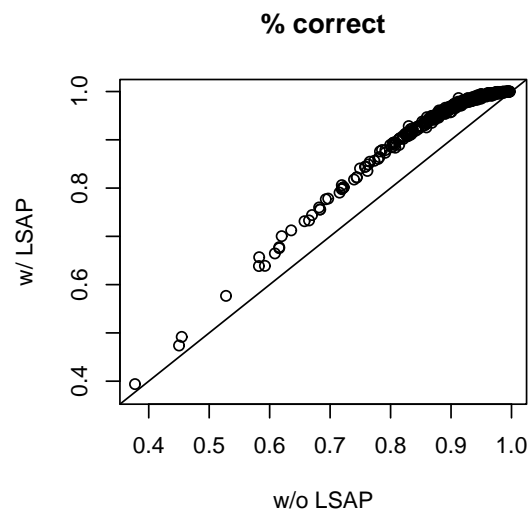
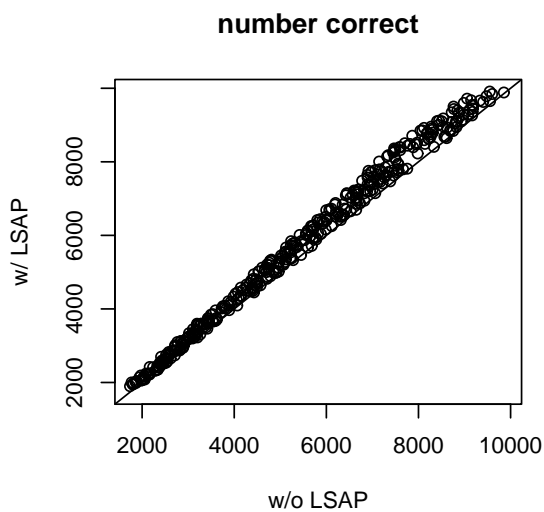
**Simulation: 400 data sets**

Criterion	Bigger w/ Constraints	Equal	Smaller w/ Constraints
$\Sigma  wt y$	266	94	40
$\Sigma_{p<0.5}(1-p)y + \Sigma_{p\geq 0.5}py$	266	94	40
$\Sigma_{p>0.9} match$	1	340	59
$\Sigma_{p>0.6} match$	0	321	79
$\%match, p > 0.9$	33	339	28
$\%match, p > 0.6$	80	331	9
$\Sigma_{p<0.1} nonmatch$	111	288	1
$\Sigma_{p<0.4} nonmatch$	100	300	0
$\%nonmatch, p < 0.1$	1	320	79
$\%nonmatch, p < 0.4$	1	316	83

## Impact of 1-1 Assignment

- Not automatic using mixture models or other procedures.
- Linear Sum Assignment Procedure: Burkard and Derigs (1980)
- Versions of the “Stable Marriage” problem

**Impact: Drastically Decrease nonmatching pairs**



## Record Linkage Hierarchical Model

Within block  $b$ , we have prior distributions:

$$P(\gamma_k = 1 | M, b) \sim \text{Beta}(\alpha_{bMk}, \beta_{bMk})$$

$$P(\gamma_k = 1 | U, b) \sim \text{Beta}(\alpha_{bUk}, \beta_{bUk})$$

$$p_{Mb} \sim \text{Beta}(\alpha_b, \beta_b)$$

and across blocks we have

$$(\alpha_{bMk}, \beta_{bMk}) \sim g(\mu_{Mk}, \Sigma_{Mk})$$

$$(\alpha_{bUk}, \beta_{bUk}) \sim g(\mu_{Uk}, \Sigma_{Uk})$$

$$(\alpha_b, \beta_b) \sim g(\mu, \Sigma).$$

The likelihood is the likelihood from the **latent class** model.

## Simulating the Posterior Distribution

Within block  $b$ ,

- Generate a “move”
  - Add a link between two unlinked records
  - Remove a link between two linked records
  - Switch two linkages

Accept or reject move.

- Draw parameters
  - same as before but “local”

◇ Draw Hyperparameters

Metropolis-Hastings step

## **Choice of Hyperprior distributions**

Look at blocks in several data sets to get idea of plausible values.

## **Advantage of the Hierarchical method**

- Power of 1-1 matching and constraints directly
- Variability across blocks is allowed
- Shrinkage:

Small blocks do not have unacceptable variability in parameter estimates

Experiments with small data sets shows this could be a concern.

## Conclusions

1. Record linkage in some census applications can be facilitated by mixture models.
2. Regression of linked results can be accomplished using estimated probabilities from record linkage.
3. Bayesian record linkage allows incorporation of prior opinion and expression of uncertainty.
4. Inequality constraints and one-to-one matching restrictions can be incorporated.
5. Ideas are being studied for Bayesian methods for allowing explicit blocking.
6. Scheuren-Winkler (1997) will be attempted within a full model of both record linkage and regression.