# COMMENTS ON HIERARCHICAL BAYESIAN RECORD LINKAGE

Michael D. Larsen, Department of Statistics, 5734 S. University Avenue, Chicago, Illinois 60637,
larsen@galton.uchicago.edu

**Key Words: Blocking, Fellegi-Sunter, Latent Class Analysis, Measurement Error, One-to-one Assignment**

## Acknowledgments

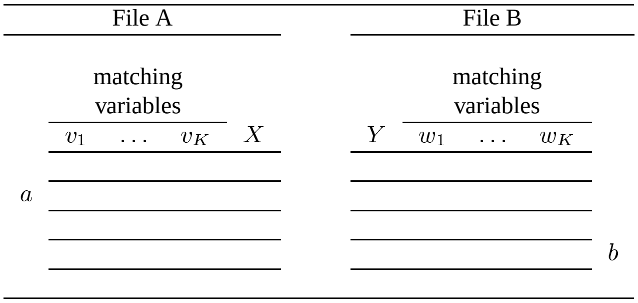
The author would like to thank the organizer of the session for the invitation to speak and the U.S. Bureau of the Census for support.

## 1 Record Linkage/File matching

A goal of record linkage is to join together two files that contain information on the same individuals, but lack unique codes to bring the pieces of information together correctly. Large-scale applications of record linkage can be found in federal statistical systems (Alvey and Jamerson 1997) and medical studies (Newcombe 1988), in which data bases are very large and processing time and accuracy are concerns.

Fellegi and Sunter (1969), formalizing ideas of Newcombe *et al.* (1959), proposed a model for record linkage. Let there be two files called file $A$ and file $B$. The set of record pairs $A \times B = \{(a,b), a \in A, b \in B\}$ is composed of two disjoint subsets: the set of true links, $M$, and the set of true nonlinks, $U$.

For each pair of records $(a, b)$, the outcomes of a1 series of comparisons are reported in a vector: $\gamma(a,b)' = \{\gamma_k(a,b), k = 1, \ldots, K\}$, where $\gamma_k(a,b) = 1$ if records $a$ and $b$ agree on comparison $k$ and 0 otherwise. If the variables in file $A$ are $v_1, \ldots, v_K$ and in file $B$ they are $w_1, \ldots, w_K$, then $\gamma_k(a,b) = 1$ if $v_k(a) = w_k(b)$. In U.S. census operations, the fields used for comparison often include first and last name, house number and street address, phone number, sex, race, age, and relation to head of household. Unique identifying numbers are typically not available or are reported with errors. The first table below illustrates the file structure. If clerks were able to review all pairs of records, then the pairs could be divided into two sets $A \times B = M \bigcup U$ (Fellegi, Sunter 1969), where $M$ contains true links and $U$ true nonlinks. The following table presents part of a simulated data set.



| 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 |

The model contains some unknown probabilities. These include the fraction of pairs that are true links (P($M$)) and nonlinks (P($U$)) and the chance of observing a pattern of responses $\gamma$ in each of the two underlying sets: P($\gamma|M$) and P($\gamma|U$). If the conditional probabilities were known, then it would be possible to a compute likelihood ratio, $\Lambda = P(\gamma|M)/P(\gamma|U)$, for each agreement-disagreement pattern $\gamma$. 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 log of the likelihood is often called a weight. The larger the weight, the more likely the pair with a given comparison vector is to be a true link.

## 2 The Use of Mixture Models for Record Linkage

Finite mixture models are useful when the population being studied is composed of two or more subpopulations that are not clearly identified (McLachlan and Peel 2000). In record linkage, before clerical review has been completed and in the absence of unique identifying information, the status of pairs as true links and true nonlinks is unknown, but real. Before clerical review is undertaken, mixture models can be applied to the agreement/disagreement patterns between record pairs in order to estimate probabilities used in calculating record linkage weights. In some applications (e.g., Larsen and Rubin 2001 and references therein), the mixture classes correspond very closely to the sets of true links and true nonlinks.

Let $g = 1, \ldots, G$ be the mixture classes. Observations that are unclassified arise from one of the $G$ classes, but class membership is unknown. The marginal probability of a comparison vector $\gamma$ is

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

Fellegi and Sunter (1969) and other authors have made an assumption of conditional independence of comparison fields within classes. The mixture model then is usually referred to as a latent class model (Goodman 1974, Haberman 1974, 1979). In such as model, the probability of comparison vector $\gamma$ in class $g$ is

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

The conditional independence assumption is not necessary and has been relaxed by Larsen and Rubin (2001) and other authors referenced therein.

In the case of binary comparisons on each field of information, the data to which the mixture models are being fit are counts in a $2^K$ table. Let record pairs be indexed $i = 1, \ldots, n$. The fact that classification of pairs into underlying classes is unknown can be represented by variables which are missing. For $i = 1, \ldots, n$ and $g = 1, \ldots, G$, let $z_{ig} = 1$ if pair $i$ is from mixture class $g$ and 0 otherwise. Then $\sum_{g=1}^{n} z_{ig} = 1$. Using Bayes' Theorem, given parameter values, it is possible to compute the probability of membership in class $g$ given a comparison vector $\gamma$:

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

It is possible also to compute the Fellegi-Sunter likelihood ratio given the mixture model parameter estimates. If some of the latent classes are associated with the true links ($g \in S_M$) and others with true nonlinks ($g \in S_U$), then

$$\sum_{g \in S_M} \pi_g P(\gamma| \text{ class g }) = P(M)P(\gamma|M)$$

and

$$\sum_{g \in S_U} \pi_g P(\gamma| \text{ class g }) = P(U)P(\gamma|U)$$

and the Fellegi-Sunter likelihood ratio is

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

It also is possible to estimate error rates, $\mu$ and $\lambda$:

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

Algorithms for maximum likelihood estimation of latent class models and of more general mixture models for cross classified tables of counts have been presented by various authors (e.g., Armstrong and Mayda 1993, Winkler 1988, 1994). Larsen and Rubin (2001) contains references to general literature and to articles specific to record linkage. Most authors use the Expectation-Maximization (EM; Dempster, Laird, and Rubin 1977) or the Expectation-Conditional Maximization (ECM; Meng and Rubin 1993) algorithms. Larsen (1994, 1996) fit Bayesian latent class and Bayesian log linear models to record linkage data. Bayesian methods for table of counts also have been presented in Gelman *et al.* (1995).

## 3 Blocking and 1-1 assignment

A few issues arise in the actual record linkage data that typically are not included in the mixture models described above. As theory for record linkage develops and modeling of data improves, it will be important to consider these dimensions of the data.

### 3.1 Blocking

A first issue is referred to as blocking. Not all record pairs $(a, b)$ are compared to one another; record pairs are compared only within blocks of records. In the census applications, the blocks are geographically defined and clearly identifiable. The blocks represent important structure in the data. It is assumed that it is unlikely that true matches exist across blocks.

The proportions of matches and nonmatches within a block varies across the blocks. For example, in the census applications, some are stable suburbs and others are highly mobile inner cities or places around universities.

The characteristics of populations in terms of the matching variables varies by block. Certain types of names are common in some areas, but rare in others. Race might not be informative in some communities, but more informative in others.

If sample sizes within blocks were not small, one could imagine applying the Fellegi-Sunter procedure separately in each block. In census geography, however, the sample sizes within blocks will not be very large.

### 3.2 One-to-one Assignment

In many applications, it is assumed that duplicates within a file are few in number or can be eliminated before the file it linked to another source. Removing duplicates might be feasible if within a file subjects use unique identification codes or records are accumulated over time (e.g., in a hospital or insurance plan) and can be unduplicated. A modification of the above procedure could also be used to locate likely duplicates.

If it is assumed that there are no duplicates in files $A$ and $B$, then there should be at most one unique match for each record in each file. That is, if $I(a,b) = 1$ if pair $(a,b)$ is a match and zero otherwise, then $\sum_{a \in block} I(a,b) \leq 1$ and $\sum_{b \in block} I(a,b) \leq 1$.

Many operations in practice today impose one-to-one matching constraints after fitting mixture models to the full set of possible pairs within blocks (Jaro 1989, 1995, Winkler 1995). A consequence of not imposing the one-to-one restrictions before estimating parameters is that the sum of expected values of indicators ($z_{ig}$'s) within a block can be larger than the size of the block or the sum over $g$ for a particular $i$ can be different from 1.

### 3.3 Logical constraints

There are other logical constraints on the latent data. Within a block, the total number of matched pairs has to be less than the smaller of the two files in that block:

$$\# \text{ linked pairs in a block } \leq \min(\#A, \#B \text{ in the block}).$$

Also, it seems reasonable to require the probability of agreeing on a field of comparison to be higher among matches than among nonmatches:

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

This constraint can be translated to a constraint about the relationships of conditional probabilities in the latent classes.

## 4 Bayesian Record Linkage

One motivation for considering a Bayesian approach to record linkage is that there is a great deal of experience with record linkage in certain settings, such as in census and health applications. There are some data sets in these areas where clerical review has verified the link/nonlink status of every pair of records (at least within blocks).

In order to implement a Bayesian mixture model approach for record linkage, one may specify a prior distributions on the probabilities of comparison patterns:

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

where the distribution are Dirichlet with parameters $\alpha_\gamma^M$ for the links and $\alpha_\gamma^U$ for the nonlinks.

In the latent class approach, one can specify independent Beta distributions for the agreement probabilities among the links and nonlinks:

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

The probability of being a true link also needs a prior distribution, which can be taken as a Beta distribution:

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

It is possible to specify the prior distribution by thinking about a table of counts for matches and for nonmatches. Information about the prior formulation can be taken from previous record linkage experiments in which clerks have reviewed the data and determined match/nonmatch status.

### 4.1 Simulation of the Posterior Distribution

The results of Bayesian record linkage could be summarized by simulating the posterior distribution of unknown parameters and of the unknown indicators of link/nonlink status $z_{ig}$. Computations of this sort can be accomplished via Data Augmentation (Tanner and Wong 1987) or Gibbs sampling (Gelfand and Smith 1990).

One difficulty in this approach, which is a problem for all similar approaches as well, is that the unknown indicators $z_{ig}$ must be generated given the current parameters. This is difficult when one wants to satisfy the constraints on the data described in the previous section.

### 4.2 An alternative Bayesian approach

Another Bayesian approach is considered by Fortini, Liseo, Nuccitelli, and Scanu (2000). These authors avoid placing prior distributions on parameters by averaging over them analytically.

These authors rely heavily on the prior distribution rather than making modeling assumptions or placing restrictions on the parameters to force their models to locate matches and nonmatches rather than some other subset of the data. Future work will compare the two approaches.

### 4.3 A Hierarchical Model

An extension of the Bayesian approach discussed in this comment is to propose a hierarchical model for record linkage. The hierarchical model could be useful in record linkage because record linkage must be performed across many sites and blocks all of which have some similarities in their parameters but some variability. The idea would be to allow each area to adapt to its own configuration of agreement/disagreement patterns, but to limit the variability by relating the parameters across sites or blocks to one another through a hierarchical model.

Within block $b$, the prior distributions on parameters could be

$$P(\gamma_k = 1|M, b) \sim \text{Beta}(\alpha_{bMk}, \beta_{bMk})$$
$$P(\gamma_k = 1|U, b) \sim \text{Beta}(\alpha_{bUk}, \beta_{bUk})$$
$$p_{Mb} \sim \text{Beta}(\alpha_b, \beta_b).$$

Across blocks the parameters of these distributions could be related to one another through the following distributions.

$$(\alpha_{bMk}, \beta_{bMk}) \sim N_2(\mu_{Mk}, \Sigma_{Mk})$$
$$(\alpha_{bUk}, \beta_{bUk}) \sim N_2(\mu_{Uk}, \Sigma_{Uk})$$
$$(\alpha_b, \beta_b) \sim N_2(\mu, \Sigma).$$

Given the number of parameters involved in these modeling specifications, the likelihood for the data could be taken as the likelihood from the latent class model.

### 4.4 Simulating the Posterior Distribution

The simulation of the posterior distribution is more involved than it was before because of the second level of the hierarchy and the number of parameters. If the prior distribution for the parameters across the blocks is chosen as above, then the simulation of the posterior distribution can be accomplished with using the Metropolis-Hastings algorithm within a Gibbs sampling sequence.

## 5 Estimation of Regression Coefficients

After file $A$ and file $B$ have been linked together, it might be of interest to analyze relationships between variables that were originally separate on the two files. If mismatch errors are introduced during record linkage , statistical analyses based on linked data can be adversely affected. Work on this problem has been done by Lahiri and Larsen (2002) and Scheuren and Winkler (1993, 1997).

Consider the following regression model

$$y_i = x_i'\beta + \epsilon_i, i = 1, \ldots, n,$$

where $x_i = (x_{i1}, \cdots, x_{ip})'$ is a vector of $p$ known covariates, $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$, and $cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j, \quad i, j = 1, \ldots n$. As in the first figure, suppose that response ($Y$) is in file $B$, the covariates ($X$) are in file $A$, and the two files are linked imperfectly. The true pairs $(x_i, y_i)$ are not observable. Instead, we observe $z_i's$ which may or may not correspond to $x_i$.

Scheuren and Winkler (1993) proposed the following model for $z_i$'s:

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

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

A naive estimator of $\beta$ would be

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

where $X = (x_1', \ldots, x_n')'$ and $Z = (z_1, \ldots, z_n)'$. This estimator is biased due to the imperfect linkage of response and predictor variables.

An improved estimator was presented by Scheuren and Winkler (1993). An iterative procedure was presented by Scheuren and Winkler (1997). Lahiri and Larsen (2002; submitted to *JASA*) developed an alternative estimator of $\beta$ and its SE. All these methods use the estimated probability of being links for adjustment of regression results.

The Bayesian approach to record linkage discussed in this paper can be used in combination with the regression adjustment procedures that use estimated probabilities in estimation.

# 6 Summary

A hierarchical Bayesian record linkage method has been outlined in general form. Important considerations, such as one-to-one matching, for modeling data in record linkage operations also have been described. The further issue analyzing data created through record linkage in subsequent analyses has been briefly presented.

The hierarchical Bayesian record linkage method has been implemented for small test data sets with very good matching information. The method will be implemented for larger data sets based on U.S. census operations. In short, Bayesian mixture models can be used to cluster pairs $(a, b)$ into links and nonlinks, as described in the Fellegi-Sunter (1969) approach, and uncertainty in error rates can be expressed through simulation.

The output the Bayesian record linkage procedure can be used in the Lahiri and Larsen (2002) method. The method of Lahiri and Larsen (2002) has been implemented using the output of
Bayesian record linkage. This works pretty well on the simulated data.

Attempts will be made to unify the record linkage model and the regression model into one fully Bayesian model. As in Scheuren and Winkler (1997), an improvement both in the record linkage and in the estimation of the regression parameters might be possible by considering the relationship between linkage probabilities and residuals in the regression.

Further work on using real prior information in the form of data from other record linkage operations will be conducted.

Finally, further research and testing will be done on efficient implementation of the sampling algorithms for Bayesian record linkage with large data sets.

# 7 References

Alvey, W., and Jamerson, B. (1997), *Record Linkage Techniques – 1997*, Proceedings of an International Workshop and Exposition. Federal Committee on Statistical Methodology, Office of Management of the Budget.

Armstrong, J .B., and Mayda, J. E. (1993), “Model-Based Estimation of Record Linkage Error Rates, ” *Survey Methodology*, 19, 137-147.

Belin, T. R., and Rubin, D. B. (1995), “A Method for Calibrating False-Match Rates in Record Linkage, ” *Journal of the American Statistical Association*, 90, 694-707.

Dempster, A. P., Laird, N. M., and Rubin, Donald B. (1977), “Maximum Likelihood from Incomplete Data Via the EM Algorithm, ” *Journal of the Royal Statistical Society*, Series B, 39, 1-22, (C/R: p22-37).

Fellegi, I., and Sunter, I. (1969), “A Theory for Record Linkage,” *Journal of the American Statistical Association*, 64, 1183-1210.

Gelfand, Alan E., and Smith, Adrian F. M. (1990), “Sampling-Based Approaches To Calculating Marginal Densities, ” *Journal of the American Statistical Association*, 85, 398-409.

Gelman, Andrew, Carlin, John B., Stern, Hal S., and Rubin, Donald B. (1995), *Bayesian Data Analysis*, Chapman & Hall.

Goodman, Leo A. (1974), “Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models, ” *Biometrika*, 61, 215-231.

Haberman, Shelby J. (1974), “Log-Linear Models For Frequency Tables Derived By Indirect Observation: Maximum Likelihood Equations, ” *The Annals of Statistics*, 2, 911-924.

Haberman, Shelby J. (1979), *Analysis of Qualitative Data: (Vol. 2), New Developments*, Academic Press.

Jaro, M. A. (1989), “Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida,” *Journal of the American Statistical Association*, 84, 414-420.

Jaro, M. A. (1995), “Probabilistic Linkage of Large Public Health Data Files, ” *Statistics in Medicine*, 14, 491-498.

Lahiri, P., and Larsen, M. D. (2002), *Regression Analysis with Linked Data*, under revision for *Journal of the American Statistical Association*.

Larsen, M.D. (1994), “Data Augmentation with Bayesian Iterative Proportional Fitting Applied to a Census Bureau Latent-Class Problem”, (1994).

Proceedings of the Government Statistics Section, American Statistical Association, pages 116-121.

Larsen, M.D. (1996), "Bayesian Approaches to Finite Mixture Models", (1996). Thesis, Harvard University, Department of Statistics.

Larsen, M. D., and Rubin, D. B. (2001), "Iterative Automated Record Linkage Using Mixture Models," *Journal of the American Statistical Association*, 96, 32-41.

McLachlan, G., and Peel, G. (2000), *Finite Mixture Models*, John Wiley & Sons.

Meng, Xiao-Li, and Rubin, Donald B. (1993), "Maximum Likelihood Estimation Via the ECM Algorithm: A General Framework, " *Biometrika*, 80, 267-278.

Neter, John, Maynes, E. Scott, and Ramanathan, R. (1965), "The effect of mismatching on the measurement of response errors," *JASA*, 60, 1005-1027.

Newcombe, Howard B. (1988), *Handbook of record linkage*: Methods for Health And Statistical Studies, Administration, and Business. Oxford University Press.

Newcombe, H.B., Kennedy, J.M., Axford, S.J., and James, A.P. (1959), "Automatic Linkage of Vital Records," *Science*, 954-959.

Scheuren, Fritz, and Winkler, William E. (1993), "Regression analysis of data files that are computer matched," *Survey Methodology*, 19, 39-58.

Scheuren, Fritz, and Winkler, William E. (1997), "Regression analysis of data files that are computer matched – Part II," *Survey Methodology*, 23, 157-165.

Tanner, Martin A., and Wong, Wing Hung. (1987), "The Calculation of Posterior Distributions By Data Augmentation," (with discussion), *Journal of the American Statistical Association*, 82, 528-540.

Winkler, W. E. (1988), "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage, " in *American Statistical Association Proceedings of Survey Research Methods Section*, pp. 667- 671.

Winkler, W. E. (1994), "Advanced Methods for Record Linkage, " in *American Statistical Association Proceedings of Survey Research Methods Section*, pp. 1994.

— (1995), "Matching and Record Linkage," in *Business Survey Methods*, ed. Cox, B. G., Binder, D. A., Chinnappa, B. N., Christianson, A., Colledge, M. J., and Kott, P. S., New York: Wiley Publications, pp. 355-384.