Advances in Record Linkage Theory: Hierarchical Bayesian Record Linkage Theory

Michael D. Larsen, Iowa State Univeristy
Department of Statistics, 220 Snedecor Hall, Ames, Iowa 50011-1210, larsen@iastate.edu

Keywords: Constraints, Fellegi-Sunter, Gibbs sampling, Hierarchical model, Latent class model, Metropolis-Hastings algorithm, Mixture model.

Abstract: In record linkage, or exact file matching, one compares two or more files on a single population for purposes of unduplication or production of an enhanced, merged database. Record linkage has many applications, including in population enumeration efforts, to create databases for epidemiological investigations, and to improve survey sample frames. Latent class and mixture models have been used to implement computerised record linkage of large databases. Probabilities that pairs of records, one record from each of two files, pertain to the same person (a match) or to different people (a nonmatch) are estimated based on model parameters and Bayes' theorem. In some settings, there is experience with similar record linkage operations that can inform prior opinions concerning model parameters. In this paper, Bayesian record linkage alternatives are developed. A hierarchical Bayesian model allows parameters to vary by file blocks, which are similar to geographical blocks in census applications. Techniques for incorporating 1-1 matching between files into the likelihood itself and computing posterior distributions of parameters and linkage indicators are presented.

1. Introduction

Record linkage (e.g., Fellegi and Sunter 1969, Newcombe et al 1959) involves comparing two or more files on the same population for purposes of unduplication of records and merging files. Record linkage is used in many applications, including population size estimation at the U.S. Bureau of the Census (Winkler 1994, 1995, and Jaro 1989, 1995), epidemiology and medical studies (Newcombe 1988, Gill 1997), sociological studies (Belin et al 2004), and survey frame improvement. See also Alvey and Jamerson (1997) and references therein.

Latent class (McCutcheon 1987) and mixture models (McLachlan and Peel 2000) have been used to model the data arising from comparing records in two files (Larsen and Rubin 2001, Winkler 1988, 1994, 1995, Jaro 1995). Although successful in many applications (Alvey and Jamerson 1997), the models used in these applications have not accounted for all restrictions in the data. In particular, forcing each record on one file to have at most a single, matching record on the other file ("1-1 matching") has been implemented post-hoc with a linear-sum assignment procedure (Burkard and Derigs 1980, Jaro 1989) to choose individual links. The 1-1 assignment procedure can effectively eliminate many candidate links that have some degree of similarity, but actually are nonlinks.

Experience from previous record linkage operations has been used informally to select models (Larsen and Rubin 2001) and restrict parameters (Winkler 1989, 1994). Bayesian approaches have been suggested by Larsen (1999a, 2002), Forinti et al (2002, 2000), and McGlinchy (2004). A new procedure will be proposed that explicitly uses the 1-1 matching assumption and allows parameter values to vary by file block.

2. Record Linkage

Suppose that there are two files, A and B, on a single population. Consider record a in file A and record b in B. Do records a and b correspond to the same person or entity? If files A and B do not have unique, accurately recorded identification numbers for every unit in both files, then it is necessary to consider the information recorded in a and b in order to answer the question. Decennial census applications at U.S. Bureau of the Census record variables including last name and first name, street number and name, age, sex, race, and relation to head of household. Files are extensively preprocessed before linkage is attempted. For example, names are standardized and coded according to Soundex codes or other scheme. Names and address fields are parsed and standardized. In the case of simple comparisons, for each pair of records (a, b) being considered, a vector of 1's and 0's indicating agreement and disagreement on K comparison fields is recorded. That is, for $a \in A$ and $b \in B$, define $\gamma(a,b) = \{\gamma(a,b)_1, \gamma(a,b)_2, \ldots, \gamma(a,b)_K\}$, where $\gamma(a,b)_k$ equals 1 (agreeement) or 0 (disagreement) on field k. Many agreements ($\gamma(a,b)$ mostly 1's) are typical of matches. Many disagreements ($\gamma(a,b)$ mostly 0's) are typical of nonmatches.

2.1. One-one Restrictions and Blocking

In some cases, it is assumed that the two data files do not contain duplicate records for any person or entity. In the case of the census, records are organized by geographical location, each household should have only one form, and efforts are made at unduplicating records. Insurance companies and medical records systems are updated continuously and efforts are made to avoid duplicate records. Record linkage could be of interest in these cases, because census follow-up operations are conducted independently in large areas and the National Death Index is matched to to existing insurance, medical, and other databases for studies such as Livingston and Ko (2005), Rauscher and Sandler (2005), and Thompson et al. (2005).

In census and other operations, the files are divided geographically into groups of records or 'blocks' that do not overlap. Blocking is used in other applications as well in order to reduce the number of record pairs being compared. It is assumed that there are no (or very few) matches across different blocks. Other operations use first letter of last name (individuals) or industry code (businesses) or state as blocking variables.

Let blocks be indexed by $s = 1, \ldots, S$. Suppose that file A has n_{a_s} records and file B has n_{b_s} records, respectively, in block s. For blocks $s = 1, \ldots, S$, $a_s = 1, \ldots, n_{a_s}$ and $b_s = 1, \ldots, n_{b_s}$, define $I(a_s, b_s) = 1$ if a and b are matches, and 0 if a and b are nonmatches. The set of match-nonmatch indicators in block s is $I_s = \{I(a_s, b_s)\}$. The 1-1 restrictions and blocking assumptions mean that $\sum_{b_s} I(a_s, b_s) \leq 1$, $\sum_{a_s} I(a_s, b_s) \leq 1$, and $\sum_{a_s} \sum_{b_{s'}} I(a_s, b_{s'}) = 0$ for $s \neq s'$. The number of matches in block s, n_{m_s} is defined and restricted under 1-1 matching as follows: $\sum_{a_s} \sum_{b_s} I(a_s, b_s) = n_{m_s} \leq \min{(n_{a_s}, n_{b_s})}$.

2.2. Prior Beliefs and Logical Relationships

Prior experience and data often are available from previous record linkage operations and sites. In previous record linkage studies, clerks at the census looked at record pairs and determined whether or not they truly were nonmatches or matches. Belin (1993, 1995), Larsen (1999b), and Larsen and Rubin (2001) found that in some census record linkage applications characteristics of populations being studied varied by area in ways that made a significant impact on estimates of parameters needed for record linkage. There were, however, consistent patterns across areas. The percentage of record pairs, one record from each of two files, under consideration that actually are matches corresponding to the same person is roughly similar across sites. The probability of agreeing on some key fields of information among matches and nonmatches are similar across sites. The probability of agreements are higher among matches than among nonmatches.

It is expected that the probability of agreeing on an individual field of comparison is higher for matches than for nonmatches: $P(\gamma_k(a,b) = 1|(a,b) \in \text{M}) > P(\gamma_k(a,b) = 1|(a,b) \in \text{U})$. Logically, the number of matches in block s, n_{m_s}, is smaller than the smaller of the number of in files A (n_{a_s}) and B (n_{b_s}). So the probability of a match in block s, p_{sM}, is less than or equal to the minimum size divided by the number of pairs: $n_{a_s} n_{b_s}$.

3. Bayesian Record Linkage Model

3.1. Bayesian Latent Class Models

The mixture model approach to record linkage models the probability of comparison vector γ using a mixture distribution:

$$\text{Pr}(\gamma) \quad = \quad \text{Pr}(\gamma|M)p_M + \text{Pr}(\gamma|U)p_U, \qquad (1)$$

where $\text{Pr}(\gamma|M)$ and $\text{Pr}(\gamma|U)$ are the probabilities of the pattern γ among the matches (M) and nonmatches (U), respectively, and p_M and $p_U = 1 - p_M$ are marginal probabilities of matches and unmatched pairs. In practice at census and Statistics Canada, models using three classes often are useful when matching individuals because estimates based on them reflect household structure (see, e.g., Larsen and Rubin 2001, Armstrong and Mayda 1993, and Winkler 1995). Databases on businesses in general would not reflect the household grouping typical of people. We will consider the situation with two classes here.

The conditional independence assumption simplifies the model by reducing the dimension within each mixture class from $2^K - 1$ parameters to K:

$$\text{Pr}(\gamma|C) \quad = \quad \prod_{k=1}^{K} \text{Pr}(\gamma_k|C)^{\gamma_k} (1 - \text{Pr}(\gamma_k|C))^{1-\gamma_k}, \qquad (2)$$

with $C \in \{M, U\}$. Interactions between comparison fields have been allowed in Larsen and Rubin (2001), Armstrong and Mayda (1993), Thibaudeau (1993), Winkler (1989), and others. Here we consider only the conditional independence model and extensions of it to a hierarchical framework.

Previous approaches have not directly enforced 1-1 linkage in the likelihood and have used the following likelihood function:

$$\prod_{s=1}^{S} \prod_{a \in A_s, b \in B_s} \text{Pr}(\gamma(a,b)), \qquad (3)$$

where $\text{Pr}(\gamma(a,b))$ is a comparison vector modeled using the mixture assumption (1). When the parameters determining $\text{Pr}(\gamma|M)$ and $\text{Pr}(\gamma|U)$ do not depend on the block from which the pairs originate and n_γ is the number of pairs of records with comparison pattern γ, the simple likelihood can be written as $\prod_{\gamma \in \Gamma} \text{Pr}(\gamma)^{n_\gamma}$.

Assuming the conditional independence model (2) and global parameters that do not vary by block, a prior distribution on parameters can be specified conveniently as the product of independent Beta distributions as follows: $p_M \sim \text{Beta}(\alpha_M, \beta_M)$, $\text{Pr}(\gamma_k(a,b) = 1|M) \sim \text{Beta}(\alpha_{Mk}, \beta_{Mk}), k = 1, \ldots, K$, and $\text{Pr}(\gamma_k(a,b) = 1|U) \sim \text{Beta}(\alpha_{Uk}, \beta_{Uk}), k = 1, \ldots, K$.

The match/nonmatch indicators $\mathbf{I} = \{I(a,b), a \in A_s, b \in B_s, s = 1, \ldots, S\}$ are unobserved. By Bayes' theorem, if the parameters were known and one does not consider restrictions from 1-1 matching, one could calculate for a pair (a,b) the probability that a and b match ($\text{Pr}(M|\gamma(a,b))$):

$$\text{Pr}(I(a,b) = 1|\gamma(a,b)) = p_M \text{Pr}(\gamma(a,b)|M)/\text{Pr}(\gamma(a,b)). \qquad (4)$$

If the match indicators $\mathbf{I}$ were known, the posterior distributions of individual parameters given values of the other parameters would be as follows: $p_M|\mathbf{I}$ has a Beta distrbution

$$\text{B}(\alpha_M + \sum_{(a,b)} I(a,b),\ \beta_M + \sum_{(a,b)} (1 - I(a,b)) \qquad (5)$$

and, for $k = 1, \ldots, K$, $\text{Pr}(\gamma_k(a,b) = 1|M, \mathbf{I}) \sim$

$$\text{B}(\alpha_{Mk} + \sum I_{ab}\gamma_k(a,b),\ \beta_{Mk} + \sum I_{ab}(1 - \gamma_k(a,b))) \qquad (6)$$

and $\text{Pr}(\gamma_k(a,b) = 1|U, \mathbf{I}) \sim$

$$\begin{aligned} &\text{B}(\alpha_{Uk} + \sum (1 - I_{ab})\gamma_k(a,b), \qquad (7) \\ &\beta_{Uk} + \sum (1 - I_{ab})(1 - \gamma_k(a,b))), \end{aligned}$$

where $I_{ab} = I(a,b)$ and sums are over all pairs allowed within the blocking structure.

The posterior distribution of parameters is simulated by sampling from alternating conditional distributions (Gibbs sampling; Geman and Geman 1984, Geland and Smith 1990) as follows.

1. Specify parameters for the prior distributions. Choose initial values of unknown parameters.
2. Repeat four steps numerous times until the distribution of draws has converged to the posterior distribution:
   (a) Draw values for the components of $\mathbf{I}$ independently from Bernoulli distributions with the probability that $I(a,b) = 1$ given by (4).
   (b) Draw a value of p_M from the distribution specified in (5) and calculate $p_U = 1 - p_M$.
   (c) Draw values of $\text{Pr}(\gamma_k(a,b) = 1|M, \mathbf{I})$ independently for $k = 1, \ldots, K$ from distributions specified in (6).
   (d) Draw values of $\text{Pr}(\gamma_k(a,b) = 1|U, \mathbf{I})$ independently for $k = 1, \ldots, K$ from distributions specified in (7).
3. Stop once the algorithm has converged.

Once the algorithm has converged, it is necessary to decide which pairs of records to designate links and nonlinks and which to leave undecided. If 1-1 restrictions are not enforced, then one could calculate the proportion of times that a record pair (a,b) has $I(a,b) = 1$ and for each record in file A assign the record in file B that has the largest proportion. If 1-1 matching is desired, the simulated probabilities (4) of matching could be supplied to a linear-sum-assignment algorithm.

There are some *restrictions on parameters* that could improve the performance of this model for record linkage. First, the range of p_M logically should be restricted to be less than or equal to the smaller of the two file sizes divied by the number of pairs under the blocking structure. Second, logically the probability of a record pair agreeing on a comparison field should be larger among matches than among nonmatches. That is, $\text{Pr}(\gamma_k|M) > \text{Pr}(\gamma_k|U)$, for $k = 1, \ldots, K$.

There are several significant limitations to this model. First, there is no explicit 1-1 matching in the likelihood (3) and without subsequent processing some records could be involved in more than one designated link. As a consequence, it was not necessary to model the number of matches overall or within individual blocks. In many applications, some records in file A and some in file B might not have any matches. One-one matching then is the assumption that records have at most one match in the other file. Second, the parameters are global and do not vary across blocks despite the fact that populations can vary greatly across blocks. Third, the conditional independence assumption was made for convenience and is not realistic. It has been relaxed in the case of maximum likelihood estimation (see Larsen and Rubin 2001 and references therein). Interactions between comparison fields within the matches and nonmatches could be allowed in the Bayesian approach as well. It is the belief of the author, however, that explicitly modeling 1-1 matching and allowing parameters to vary by block will be more beneficial than modeling interactions globally.

3.2. A Hierachical Bayesian Model

A hierarchical model for record linkage will specify distributions of parameters within blocks $s = 1, \ldots, S$. The likelihood used in this section is given by likelihood (3) with parameters varying by block. The probabilities of agreeing on fields of information are allowed to vary by block as follows $p_{sMk} = \text{Pr}(\gamma_k = 1|M, s) \sim \text{Beta}(\alpha_{sMk}, \beta_{sMk})$ and $p_{sUk} = \text{Pr}(\gamma_k = 1|U, s) \sim \text{Beta}(\alpha_{sUk}, \beta_{sUk})$ independently across blocks, fields, and classes (M and U). The restriction that $p_{sMk} \geq p_{sUk}$ will be assumed.

Hyperpriors distributions are placed on transformed versions of the Beta parameters. The distributions are independent across blocks, fields, and groups. These transformations appeared in Larsen (2004): $\theta_{sMk} = \text{logit}(\alpha_{sMk}/(\alpha_{sMk} + \beta_{sMk})) \sim N(\mu_{\theta Mk}, \sigma^2_{\theta Mk})$, $\theta_{sUk} = \text{logit}(\alpha_{sUk}(\alpha_{sUk} + \beta_{sUk})) \sim N(\mu_{\theta Uk}, \sigma^2_{\theta Uk})$, $\tau_{sMk} = \log(\alpha_{sMk} + \beta_{sMk}) \sim N(\mu_{\tau Mk}, \sigma^2_{\tau Mk})$, and $\tau_{sUk} = \log(\alpha_{sUk} + \beta_{sUk}) \sim N(\mu_{\tau Uk}, \sigma^2_{\tau Uk})$. Note that there is a unique bivariate inverse transformation: $\alpha_{sCk} = e^{\tau_{sCk}}\text{logit}^{-1}(\theta_{sCk})$ and $\beta_{sCk} = e^{\tau_{sCk}}\text{logit}^{-1}(1 - \theta_{sCk})$ for $C = M, U$. The restriction noted in the previous paragraph *does not* mean that, for $k = 1, \ldots, K$, $\theta_{sMk} \geq \theta_{sUk}$; the restriction only constrains the parameters p_{sMk} and p_{sUk}. It would be possible to use a prior distribution with the constraint that $\theta_{sMk} \geq \theta_{sUk}$ as well.

The probability of belonging to class M in block s, p_{sM}, is given a Beta$(\alpha_{sM}, \beta_{sM})$ prior distribution. The hyperprior distrbutions are $\theta_{sM} = \text{logit}(\alpha_{sM}/(\alpha_{sM} + \beta_{sM})) \sim N(\mu_{\theta M}, \sigma^2_{\theta M})$ and $\tau_{sM} = \log(\alpha_{sM} + \beta_{sM}) \sim N(\mu_{\tau M}, \sigma^2_{\tau M})$, and are independent of the other hyperpriors. The restriction that p_{sM} is smaller than the minimum of n_{A_s} and n_{B_s} divided by the number of pairs $n_{A_s}n_{B_s}$ is enforced in this model. If it were not, the small sample size and great variability across blocks would surely produce poor results for some blocks. Note that $\alpha_{sM} = e^{\tau_{sM}}\text{logit}^{-1}(\theta_{sM})$ and $\beta_{sM} = e^{\tau_{sM}}\text{logit}^{-1}(1 - \theta_{sM})$.

3.3. Simulating the Posterior Distribution

The posterior distribution of parameters and unobserved match/nonmatch indicators will be simulated using Gibbs sampling. The conditional distributions for the hyperparameters will be sampled using the Metropolis-Hastings (MH) algorithm (Hastings 1970) within the Gibbs sampling framework. The procedure iterates through draws of full conditional distributions:

1. Choose hyperparameter distributions. That is, specify $(\mu_{\theta M}, \sigma^2_{\theta M})$ and, for $k = 1, \ldots, K$, specify $(\mu_{\theta Mk}, \sigma^2_{\theta Mk})$, $(\mu_{\theta Uk}, \sigma^2_{\theta Uk})$, $(\mu_{\tau Mk}, \sigma^2_{\tau Mk})$, and $(\mu_{\tau Uk}, \sigma^2_{\tau Uk})$.

2. Generate initial values of $(\alpha_{sM}, \beta_{sM})$ and, for $k = 1, \ldots, K$, $(\alpha_{sMk}, \beta_{sMk})$, $(\alpha_{sUk}, \beta_{sUk})$ from their prior distributions.

3. Assign an initial match/nonmatch configuration **I**. Since 1-1 matching is not being forced, but constraints on the parameters and proportion of matches are, the algorithm of Section 3.1 with analogous parameter constraints could be run for several iterations.

4. Cycle through the following steps numerous times until convergence. Let I_{ab} denote $I(a, b)$.

   (a) For $s = 1, \ldots, S$, draw p_{sM} from its conditional distribution given the current indicators $\mathbf{I}_s$ and values of $(\alpha_{sM}, \beta_{sM})$. Specifically, $p_{sM}|I_s, \alpha_{sM}, \beta_{sM} \sim$ Beta$(\alpha_{sM} + \sum I_{ab}, \beta_{sM} + n_{a_s}n_{b_s} - \sum I_{ab})$, where the sum is over all pairs (a, b) in block s. Enforce the constraint: $p_{sM} \leq$ min$(n_{a_s}, n_{b_s})/(n_{a_s}n_{b_s})$.

   (b) For $s = 1, \ldots, S$, $k = 1, \ldots, K$, draw p_{sMk} and p_{sUk} from their conditional distribution given the current indicators $\mathbf{I}_s$, the comparison vectors γ_s in block s, and values of $(\alpha_{sCk}, \beta_{sCk}), C \in \{M, U\}$. Specifically, $p_{sMk}|I_s, \gamma_s, \alpha_{sMk}, \beta_{sMk} \sim$ Beta$(\alpha_{sMk} + \sum_s I_{ab}\gamma_k(a, b), \beta_{sMk} + \sum_s I_{ab}(1 - \gamma_k(a, b)))$, $p_{sUk}|I_s, \gamma_s, \alpha_{sUk}, \beta_{sUk} \sim$ Beta$(\alpha_{sUk} + \sum_s(1 - I_{ab})\gamma_k(a, b), \beta_{sUk} + \sum_s(1 - I_{ab})(1 - \gamma_k(a, b)))$, and $p_{sMk} \geq p_{sUk}$, where sums are over all pairs (a, b) in block s.

   (c) For $s = 1, \ldots, S$, use the MH algorithm (Hastings 1970; see also Gelman 1992 and Gelman et al 2004, chapter 11) to draw values of hyperparameters θ_{sM} and τ_{sM} from their full conditional distributions. See Appendix A for details of this and the next two steps.

   (d) For $s = 1, \ldots, S$, $k = 1, \ldots, K$, use the MH algorithm to draw values of hyperparameters θ_{sMk} and τ_{sMk}.

   (e) For $s = 1, \ldots, S$, $k = 1, \ldots, K$, use the MH algorithm to draw values of hyperparameters θ_{sUk} and τ_{sUk}.

   (f) For $s = 1, \ldots, S$, $a = 1, \ldots, n_{a_s}$, and $b = 1, \ldots, n_{b_s}$, given values of p_{sM} and, for $k = 1, \ldots, K$, p_{sMk} and p_{sUk}, draw a value of $I(a, b)$ from a Bernoulli distribution with the following probability: $p_{sM}\prod_{k=1}^{K}\left[p_{sMk}^{\gamma_k(a,b)}(1 - p_{sMk})^{1-\gamma_k(a,b)}\right]$/den, where den $= p_{sM}\prod_{k=1}^{K}\left[p_{sMk}^{\gamma_k(a,b)}(1 - p_{sMk})^{1-\gamma_k(a,b)}\right] + (1 - p_{sM})\prod_{k=1}^{K}\left[p_{sUk}^{\gamma_k(a,b)}(1 - p_{sUk})^{1-\gamma_k(a,b)}\right]$.

5. Stop once the algorithm has converged.

Note that 1-1 restrictions are not imposed on the **I** matrix. The size of the candidate match class in each block is controlled in 4(a) by keeping p_{sM} small. Once the algorithm has converged, it is necessary to decide which pairs of records to designate links and nonlinks and which to send to clerical review or leave undecided. Suggestions were made at the end of Section 3.1. Metropolis-Hastings and algorithm details are in Appendix A.

3.4. Hierachical Bayesian 1-1 Model

In this section, the 1-1 linkage assumption will be enforced in the set of indicators **I**. The hierarchical specification of Section 3.2 will continue to be used. In order to use the non-hierarchical model with 1-1 restrictions, one would need to combine the appropriate modeling assumptions and prior distributions from Section 3.1 and this section.

Define n_{m_s} to be the number of matches in block s, $s = 1, \ldots, S$. By definition, $n_{m_s} \leq \min(n_{a_s}, n_{b_s})$. The prior distribution for n_{m_s}, independently for each s, is taken to be

$$n_{m_s} \sim \text{Binomial}(\min(n_{a_s}, n_{b_s}), p_s), \tag{8}$$

where $p_s \sim$ Beta(α_p, β_p). If $\alpha_p = 4$ and $\beta_p = 1$, then $Ep_s = 0.8$, $SDp_s = 0.16$, and the distribution is skewed strongly left. If $\alpha_p = 4.5$ and $\beta_p = 1.5$, then $Ep_s = 0.75$, $SDp_s = 0.16$, and the distribution is skewed left, but not quite so strongly. The parameters p_{sM} do not play a role in this model.

Let $I_s = \{I(a, b), a \in A_s, b \in B_s\}$ for $s = 1, \ldots, S$. The prior distribution for I_s is taken to be uniform on the space of possible matching configurations:

$$P(I_s|n_{m_s}) = \left[\binom{n_{a_s}}{n_{m_s}}\binom{n_{b_s}}{n_{m_s}}n_{m_s}!\right]^{-1}. \tag{9}$$

Without examining the data to some degree, it would not be possible to assign another prior distribution. In the census application, it would be reasonable if records are grouped by household to place higher probability on records in the same household within blocks.

Given values for I, the likelihood for parameters is Pr$(\gamma|I)$:

$$\prod_{s=1}^{S}\left[\prod_{a\in A_s,b\in B_s}\left(\prod_{k=1}^{K}p_{sMk}^{\gamma_k(a,b)}(1-p_{sMk})^{1-\gamma_k(a,b)}\right)^{I(a,b)}\left(\prod_{k=1}^{K}p_{sUk}^{\gamma_k(a,b)}(1-p_{sUk})^{1-\gamma_k(a,b)}\right)^{1-I(a,b)}\right]$$

$$=\prod_{s=1}^{S}\left[\prod_{a\in A_s,b\in B_s,(a,b)\in M}\prod_{k=1}^{K}p_{sMk}^{\gamma_k(a,b)}(1-p_{sMk})^{1-\gamma_k(a,b)}\prod_{a\in A_s,b\in B_s,(a,b)\in U}\prod_{k=1}^{K}p_{sUk}^{\gamma_k(a,b)}(1-p_{sUk})^{1-\gamma_k(a,b)}\right] \tag{10}$$

As mentioned before, the parameters p_{sM} are not used in this model. Let the prior distributions for p_{sMk} and p_{sUk}, $s = 1, \ldots, S$, $k = 1, \ldots, K$ and their associated hyperprior distributions be the same as in Section 3.2.

3.5. Simulating the 1-1 Posterior Distribution

The posterior distribution of parameters and unobserved match/nonmatch indicators will be simulated using Gibbs sampling with Metropolis-Hastings (MH) steps. The procedure iterates through draws of full conditional distributions.

1. Choose hyperparameter distributions by specifying α_p and β_p and, for $k = 1, \ldots, K$, $(\mu_{\theta Mk}, \sigma^2_{\theta Mk})$, $(\mu_{\theta Uk}, \sigma^2_{\theta Uk})$, $(\mu_{\tau Mk}, \sigma^2_{\tau Mk})$, and $(\mu_{\tau Uk}, \sigma^2_{\tau Uk})$.

2. Generate initial values in blocks $s = 1, \ldots, S$ for matching variables $k = 1, \ldots, K$ of $(\alpha_{sMk}, \beta_{sMk})$ and $(\alpha_{sUk}, \beta_{sUk})$ from their prior distributions.

3. Assign an initial match/nonmatch configuration **I**. Since 1-1 matching is being forced, the algorithms of Sections 3.1 and 3.2 with appropriate constraints on parameters followed by a linear sum assignment procedure (Burkard and Derigs 1980) could be used to produce an initial **I**. In block s, $n_{m_s} = \sum_{a \in A_s} \sum_{b \in B_s} I(a, b)$.

4. Cycle through (a)-(e) until the distribution of drawn values converges to the target posterior distribution.

   (a) For $s = 1, \ldots, S$, draw p_s from its conditional distribution given the current indicators $\mathbf{I}_s$ (and hence n_{m_s}) and values of (α_p, β_p). Specifically, $p_s | I_s, \alpha_p, \beta_p \sim$ Beta$(\alpha_p + n_{m_s}, \beta_p + \min(n_{a_s}, n_{b_s}) - n_{m_s})$.

   (b) For $s = 1, \ldots, S$ and $k = 1, \ldots, K$ draw p_{sMk} and p_{sUk} from their conditional distribution as described in step (4b) of Section 3.3.

   (c) For $s = 1, \ldots, S$ and $k = 1, \ldots, K$, use the MH algorithm to draw values of hyperparameters θ_{sMk} and τ_{sMk} as described in Appendix A step (d).

   (d) For $s = 1, \ldots, S$ and $k = 1, \ldots, K$, use the MH algorithm to draw values of hyperparameters θ_{sUk} and τ_{sUk} as described in Appendix A step (e).

   (e) For $s = 1, \ldots, S$, use the MH algorithm to draw values of $\mathbf{I}_s$ and n_{m_s} from their full conditional distributions. See Appendix B for details of this step.

5. Stop once the algorithm has converged.

Note that 1-1 restrictions are imposed on the **I** matrix. The size of the match class in block s is explicitly controlled by the fact that $n_{m_s} \leq \min(n_{a_s}, n_{b_s})$; $0 < p_s < 1$. Once the algorithm has converged, it is necessary to decide which pairs of records to designate as links and nonlinks.

4. Conclusions and Future Work

A novel hierarchical Bayesian model for record linkage has been presented and implemented. The model allows probabilities to vary by block and reflect local information. 1-1 matching restrictions are imposed in the likelihood. Indicators of match status are sampled using Gibbs sampling and Metropolis-Hastings.

It will be interesting to apply these methods to data from census, NCHS, and other sources. An automated system for applying these models to new sets of files would be useful in this regard. In a real application, one could consider better specifications of prior distributions for the record linkage model parameters and the use of training data. In some applications, the size of the files will be a challenge. In order to speed computations, one might consider parallel computations by, for example, block.

The algorithm's perfomance could be improved by studying tuning parameters and the order of sampling cycles within Metropolis-Hastings (MH) and Gibbs algorithms. One could study the sensitivity of results to the specification of hyperprior distributions. If some MH draws for some parameters and elements of **I** infrequently lead to changes in the values, then one could examine methods for increaseing the frequency of accepting MH moves. In particular, one could consider combining two or more attempted moves into one step.

Two extensions related to the record linkage model can be studied. First, one can consider expanded definitions of the agreement/disagreement comparisons for the matching variables. That is, one could allow partial agreement, missing values, and string comparator metrics (Winkler 1993, 1994). Second, in some applications, one could consider more fully using household structure. In some applications at census, household structure is reflected in part by the use of three latent classes in the mixture model (Larsen and Rubin 1999 and references therein).

Another direction for development in the future is the Bayesian analysis of files that are created through record linkage operations. Lahiri and Larsen (2005) extended Scheuren and Winkler (1993) on adjusting for the bias that arises due to errors in matching. One could imagine a feed-back loop, as in Scheuren and Winkler (1997), where points with large residuals in a linear regression model are more likely than their agreement patterns alone suggest to be nonmatches and points that are very certain to be matches have more influence on a linear regression fit.

Acknowledgement

This work has been supported in part by the U.S. Bureau of the Census, the National Center for Health Statistics, and a subcontract to NSF grant EIA-0131884 awarded to NISS. The author is solely responsible for the content.

References

Alvey, W., and Jamerson, B. (1997), *Record Linkage Techniques – 1997*, Proc. of an International Workshop and Exposition. Federal Committee on Statistical Methodology, OMB.

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

Belin, T. R. (1993). Evaluation of sources of variation in record linkage through a factorial experiment. *Surv. Meth.* **19**, 13-29.

Belin, T. R., Ishwaran, H., Duan, N., Berry, S., and Kanouse, D. (2004). Identifying likely duplicates by record linkage in a survey of prostitutes. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. Gelman, A., and Meng, X. L., editors. New York: Wiley.

Belin, T. R., and Rubin, D. B. (1995). A method for calibrating false match rates in record linkage. *JASA*, 90, 694- 707.

Burkard, R.E., and Derigs, U. (1980). Assignment and Matching Problems: Solution Methods with FORTRAN-Programs. *Lecture Notes in Economics and Mathematical Systems, No. 184*, Springer-Verlag: Berlin, Heidelberg, New York, pp. 1-11.

Fellegi, I. P., and Sunter, A. B. (1969). A Theory for Record Linkage. *JASA*, 64, 1183-1210.

Fortini, M., Liseo, B., Nuccitelli, A., and Scanu, M. (2000). On Bayesian Record Linkage, *Bayesian Methods with Applications to Science, Policy, and Official Statistics*: Selected Papers from ISBA 2000: The Sixth World Meeting of the International Society for Bayesian Analysis. Editor E. I. George, 155-164.

Fortini, M., Nuccitelli, A., Liseo, B., and Scanu, M. (2002). Modelling issues in record linkage: A Bayesian perspective. *Proc. of the ASA, Survey Research Methods Section*, 1008-1013.

Gelfand, A.E., Smith, Adrian F.M. (1990). Sampling-based approaches to calculating marginal densities, *JASA*, 85, 398-409

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*, 2nd edition. Chapman & Hall/CRC.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Mach. Intel.*, 6, 721-741

Gill, L. E. (1997). OX-LINK: The Oxford Medical Record Linkage System Demonstration of the PC Version. *Record Linkage Techniques – 1997*, Federal Committee on Statistical Methodology, Office of Management of the Budget, 491.

Hastings, W. K. (1970). Monte Carlo sapling methods using Markov chains and their applications, *Biometrika*, 57, 97-109.

Jaro, M. A. (1989). Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *JASA*, 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. (2005). Regression Analysis with Linked Data. *JASA*, 100, 222-230.

Larsen, M.D. (1999a). Multiple Imputation Analysis of Records Linkage Using Mixture Models. *Proc. of the Statistical Society of Canada, Survey Methods Section*, 65-71.

Larsen, M.D. (1999b). Predicting the Residency Status for Administrative Records that Do Not Match Census Records. *Administrative Records Research Memorandum Series*, #20, Bureau of the Census, U.S. Department of Commerce.

Larsen, M.D. (2002). Comment on Hierarchical Bayesian Record Linkage. *Proc. of the ASA, Section on Bayesian Statistical Science*, CDROM: 1995-2000.

Larsen, M.D. (2004), Record linkage using finite mixture models, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. Gelman, A., and Meng, X. L., editors. New York: Wiley, 309-318

Larsen, M.D. (2005). Hierarchical Bayesian Record Linkage Theory. ISU Preprint #05-3. Dept. of Statistics, Iowa State Univ.

Larsen, M. D., and Rubin, D. B. (2001). Iterative automated record linkage using mixture models. *JASA*, 96, 32-41.

Livingston, E. H., and Ko, C. Y. (2005). Effect of diabetes and hypertension on obesity-related mortality. *Surgery*, 137 (1): 16-25.

McCutcheon, A. L. (1987). *Latent class analysis*. Sage Publications, Inc.: Newbury Park, CA; London.

McGlinchy, M. (2004). A Bayesian record linkage methodology for multiple imputation of missing links. *Proc. of the ASA, Section on Survey Research Methods*. Alexandria, VA: CDROM.

Newcombe, H. B. (1988), *Handbook of record linkage: Methods for health and statistical studies, administration, and business*, Oxford University Press: Oxford.

Newcombe, H.B., Kennedy, J.M., Axford, S.J., and James, A.P. (1959). Automatic Linkage of Vital Records. *Science*, 954-9.

Rauscher, G. H., and Sandler, D. P. (2005). Validating cancer histories in deceased relatives. *Epidemiology*, 16 (2): 262-265.

Scheuren, F., and Winkler, W.E. (1993). Regression analysis of data files that are computer matched. *Surv. Meth.*, 19, 39-58.

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

Thibaudeau, Y. (1993). The Discrimination Power of Dependency Structures in Record Linkage. *Surv. Meth.*, 19, 31-38.

Thompson, D., Kriebel, D., Quinn, M. M., Wegman, D. H., and Eisen, E. A. (2005). Occupational exposure to metalworking fluids and risk of breast cancer among female autoworkers. *American Journal of Industrial Medicine*, 47 (2): 153-160.

Winkler, W. E. (1988). Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage. *ASA Proc. of Survey Research Methods Section,* 667- 671.

Winkler, W. E. (1989). Near automatic weight computation in the Fellegi-Sunter model of record linkage. *Proc. of the Bureau of the Census Annual Research Conference,* 5, 145-155.

Winkler, W. E. (1993). Improved decision rules in the Fellegi-Sunter model of record linkage. *ASA Proc. of Survey Research Methods Section*, 274-279.

Winkler, W. E. (1994). Advanced Methods for Record Linkage. *ASA Proc. of Survey Research Methods Section*, 467-472.

Winkler, W. E. (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, 355-384.

Appendix A. MH for the Hierarchical Model

Details of the three Metropolis-Hastings (Hastings 1970) steps in the simulation procedure of Section 3.3 are presented below.

(c). For $s = 1, \ldots, S$, use the Metropolis-Hastings algorithm (Hastings 1970; see also Gelman 1992 and Gelman et al 2004 chapter 11) to draw values of hyperparameters θ_{sM} and τ_{sM} from their full conditional distributions. Specifically, given current values of θ_{sM} and τ_{sM} (and hence α_{sM} and β_{sM}), $\mathbf{I}_s$, and other parameters,

(i) Define tuning constants $h_{\theta M} > 0$ and $h_{\tau M} > 0$.

(ii) Draw $u \sim$ Uniform$(0, 1)$, $\theta^* \sim N(\theta_{sM}, \sigma^2_{\theta M}/h_{\theta M})$, and $\tau^* \sim N(\tau_{sM}, \sigma^2_{\tau M}/h_{\tau M})$.

(iii) Calculate $\alpha^* = e^{\tau^*}\text{logit}^{-1}(\theta^*)$ and $\beta^* = e^{\tau^*}\text{logit}^{-1}(1 - \theta^*)$.

(iv) Calculate r as the minimum of 1 and $p_{sM}^{\alpha^* - \alpha_{sM}}(1 - p_{sM})^{\beta^* - \beta_{sM}} \times \exp\left(-\frac{h_{\theta M}}{\sigma^2_{\theta M}}(\theta_{sM} - \theta^*)^2\right) \exp\left(-\frac{h_{\tau M}}{\sigma^2_{\tau M}}(\tau_{sM} - \tau^*)^2\right)$.

(v) If $u \leq r$, let $\theta_{sM} = \theta^*$ and $\tau_{sM} = \tau^*$.
Otherwise, let θ_{sM} and τ_{sM} remain the same.

(d). For $s = 1, \ldots, S$ and $k = 1, \ldots, K$, use the Metropolis-Hastings algorithm to draw values of hyperparameters θ_{sMk} and τ_{sMk}. Specifically, given current values of θ_{sMk} and τ_{sMk} (and hence α_{sMk} and β_{sMk}), $\mathbf{I}_s$, and other parameters, follow the steps outlined in step (c) above but with all M indexes replaced by Mk's.

(e). For $s = 1, \ldots, S$ and $k = 1, \ldots, K$, use the Metropolis-Hastings algorithm to draw values of hyperparameters θ_{sUk} and τ_{sUk}. Specifically, given current values of θ_{sUk} and τ_{sUk} (and hence α_{sUk} and β_{sUk}), $\mathbf{I}_s$, and other parameters, follow the steps outlined in step (c) above but with all M indexes replaced by Uk's.

The tuning parameters $h_{\theta M}$ and $h_{\tau M}$ are chosen so that the drawn values of the parameters are accepted approximately 23-44% of the time (Gelman et al. 2004 chapter 11.9). Thus the algorithm could be run for several iterations to assess the acceptance rate, adapting the tuning paramters as necessary. A second phase then could be initiated with fixed values.

Appendix B. Hierarchical 1-1 Model MH Steps

Here the updating step for the number of matches, n_{m_s}, and the configuration of matches and nonmatches, $\mathbf{I}_s$, for blocks $s = 1, \ldots, S$ is described. It is assumed that current values of parameters and hyperparameters are given. Each block is updated separately. Given the value of a match/nonmatch configuration $\mathbf{I}_s$, the unknown parameters of the model are drawn as described in Section 3.5.

Let $\gamma_s = \{\gamma(a, b), a \in A_s, b \in B_s\}$ be the collection of comparison vectors for all pairs in block s. For notational convenience, let $\alpha_s = (\alpha_{sMk}, \alpha_{sUk})$, $\beta_s = (\beta_{sMk}, \beta_{sUk})$, $\mu = (\mu_{\theta Mk}, \mu_{\theta Uk}, \mu_{\tau Mk}, \mu_{\tau Uk})$, and $\sigma^2 = (\sigma^2_{\theta Mk}, \sigma^2_{\theta Uk}, \sigma^2_{\tau Mk}, \sigma^2_{\tau Uk})$ ($k = 1, \ldots, K$ in each case) be collections of hyperparameters. For block s, the full conditional distribution of $(n_{m_s}, \mathbf{I}_s, \gamma_s)$ is Pr$(n_{m_s}, \mathbf{I}_s, \gamma_s | \{p_{sMk}, p_{sUk}, k = 1, K\}, p_s, \alpha_s, \beta_s, \mu, \sigma^2)$, which equals

$$\text{Pr}(n_{m_s}|p_s)\text{Pr}(\mathbf{I}_s|n_{m_s})\text{Pr}(\gamma_s|\mathbf{I}_s, \{p_{sMk}, p_{sUk}, k = 1, K\}), \quad (11)$$

which is non-zero if and only if the 1-1 and match class size restrictions of Section 2.1 are fulfilled. The distributions listed in (11) are discrete.

Here we propose incremental ways of modifying n_{m_s} and I_s to cover the space of possible configurations and to produce higher probabilities of change across iterations. Three basic "moves" or modifications of n_{m_s} and I_s will be considered. First, one matching pair can be turned into a nonmatching pair: $n^*_{m_s} = n_{m_s} - 1$ and $I(a, b)$ changes from one to zero for some (a, b) in block s. Second, one nonmatching pair is grouped with the matches: $n^*_{m_s} = n_{m_s} + 1$ and $I(a, b)$ changes from zero to one for some (a, b) such that, before changing the indicator to one, $\sum_{a \in A_s} I(a, b) = 0$ and $\sum_{b \in B_s} I(a, b) = 0$. Third, $n^*_{m_s} = n_{m_s}$ is unchanged, but I^*_s is different from I_s. The changes in I_s that will be considered will involve at most two records from A_s and two from B_s. Ideas behind such moves are described here; see Larsen (2005) for more details.

B.1. Move 1: $n^*_{m_s} = n_{m_s} - 1$

In this movement, one pair currently designated to be a match is changed to a nonmatch designation. One option chooses a matched pair from block s with uniform probability. This option likely is not too efficient. Option 2 chooses a matched pair based on the probability that the pair is a nonmatch given that one among the matches is a nonmatch.

That is, pick a matched pair (a_i, b_j) at random with the probability of dropping pair (a_i, b_j) as given in Larsen (2005).

Given that n_{m_s} in some blocks might not be too large, the computation of probabilities above in some applications might be reasonable. Pairs of records that agree on all or almost all comparisons and that have low levels of agreement with other potential matches likely would not be selected to be dropped. Pairs of records that have more disagreements and that have alternative matches should be dropped more readily.

As for option 1, the inverse move is to add the deleted pair of records to the set of designated matches (see Move 2 below). Let $\Pr(\text{drop pair}(a_i, b_j))$ be the probability of dropping pair (a_i, b_j) from the match set. Let $\Pr(\text{add pair}(a_i, b_j))$ be the probability of adding pair (a_i, b_j). The acceptance probability for the MH algorithm is the minimum value of 1 and

$$\frac{\Pr(n^*_{m_s}, \mathbf{I}^*_s|\text{param. values}) \Pr(\text{add pair}(a_i, b_j))}{\Pr(n_{m_s}, \mathbf{I}_s|\text{param. values}) \Pr(\text{drop pair}(a_i, b_j))}.$$

B.2. Move 2: $n^*_{m_s} = n_{m_s} + 1$

In this movement, one pair currently designated to be a nonmatch is changed to a match designation. A first option chooses a nonmatched pair from block s with uniform probability. Such an approach also is not likely to be efficient. Option 2 chooses a nonmatched pair based on the probability that the pair is a match given that one among the nonmatches is a match. The probability of adding pair (a_i, b_j) is given in Larsen (2005).

Pairs of records that disagree on all or almost all comparisons are not likely to be added. Pairs of records that are current nonmatches but agree on many fields are likely to be added. As for option 1, the inverse move is to delete the added pair of records from the set of designated nonmatches (see Move 1 above). The acceptance r value for the MH algorithm is

$$\frac{\Pr(n^*_{m_s}, \mathbf{I}^*_s|\text{param. values}) \Pr(\text{drop pair}(a_i, b_j))}{\Pr(n_{m_s}, \mathbf{I}_s|\text{param. values}) \Pr(\text{add pair}(a_i, b_j))}.$$

B.3. Move 3: n_{m_s} unchanged but I_s altered

In this movement, three things can happen: two matches can switch pairs, a matched pair can replace one of its units with an unmatched pair, or a matched pair can be dropped and replaced with another matched pair.

Variation 1: Two matches switch pairings
Randomly select two matched pairs, (a_i, b_j) and (a_k, b_l), with probability $2/(n_{m_s}(n_{m_s} - 1))$ and switch the pairings: (a_i, b_l) and (a_k, b_j). That is, change $I(a_i, b_j)$ and $I(a_k, b_l)$ from one to zero and $I(a_i, b_l)$ and $I(a_k, b_j)$ from zero to one. The reverse move is to undo the switch. The acceptance probability of the MH algorithm is the minimum of one and $(P(\gamma_{il}|M, s)P(\gamma_{kj}|M, s)P(\gamma_{ij}|U, s)P(\gamma_{kl}|U, s)) / (P(\gamma_{ij}|M, s)P(\gamma_{kl}|M, s)P(\gamma_{il}|U, s)P(\gamma_{kj}|U, s))$.

It would be possible to select two matched pairs with non uniform probabilities, but doing so could be computationally expensive (see Larsen 2005). A less computationally intense approach would randomly choose one matched pair, say (a_i, b_j), with uniform probability ($1/n_{m_s}$) and a second matched pair with nonuniform probability. Given that pair (a_i, b_j) is going to be broken and switched with another pair from the current matches, one could select the pair (a_k, b_l) with probability

$$\frac{P(\gamma_{il}|M, s)P(\gamma_{kj}|M, s)P(\gamma_{ij}|U, s)P(\gamma_{kl}|U, s)}{\sum_{(k',l')\neq(i,j)} P(\gamma_{il'}|M, s)P(\gamma_{k'j}|M, s)P(\gamma_{ij}|U, s)P(\gamma_{k'l'}|U, s)}.$$

If a similar reverse move is considered, then the MH r value is

$$\frac{\sum_{(k',l')\neq(i,j)} P(\gamma_{il'}|M, s)P(\gamma_{k'j}|M, s)P(\gamma_{ij}|U, s)P(\gamma_{k'l'}|U, s)}{\sum_{(i'l')\neq(k,j)} P(\gamma_{i'j}|M, s)P(\gamma_{kl'}|M, s)P(\gamma_{i'l'}|U, s)P(\gamma_{kj}|U, s)}.$$

Variation 2: A matched pair replaces one of its matching records with a nonmatching record
In this move, a matched pair of records is randomly chosen and one of its component records is replaced with a record from the same file in the same block that does not have a designated match. That is, suppose $I(a_i, b_j) = 1$ and the matched pair (a_i, b_j) is chosen. One of the matched pairs can be chosen with uniform probability: $1/n_{m_s}$. A record a_k in file A without a match satisfies $\sum_{j'} I(a_k, b_{j'}) = 0$. A record b_l in file B without a match satisfies $\sum_{i'} I(a_{i'}, b_l) = 0$. There are $n_{a_s} + n_{b_s} - 2n_{m_s}$ nonmatched records in block s. One option is to choose a nonmatched record randomly. The reverse move would involve switching to the initial pairings. If the A-record a_i is replaced through random selection with A-record a_k, the MH acceptance probability is the minimum of one and $P(\gamma_{kj}|M, s)P(\gamma_{ij}|U, s)/\left(P(\gamma_{ij}|M, s)P(\gamma_{kj}|U, s)\right)$. If the B-record b_j is replaced through random selection with B-record b_l, the MH acceptance probability is the minimum of one and $P(\gamma_{il}|M, s)P(\gamma_{ij}|U, s)/\left(P(\gamma_{ij}|M, s)P(\gamma_{il}|U, s)\right)$.

Another way to choose the replacement record is to compute the probability given current parameter values that a particular nonmatching record is a match, assuming that pair (a_i, b_j) is a nonmatching pair. See Larsen (2005) for details.

Var. 3: Delete a matched pair; Pair 2 unmatched records
The last move that will be contemplated is the deletion of a matched pair and the joining of two unmatched records. If (a_i, b_j) is a match and a_k and b_l are unmatched records, the move entails setting $I(a_i, b_j) = 0$ and $I(a_k, b_l) = 1$. This is in effect almost the combination of the first two moves: removal of a match and addition of a new match other than the one that was removed. An acceptance probability for the MH algorithm can be computed as the product of appropriately modified probabilities associated with Moves 1 and 2.