EUROPEAN JOURNAL OF PURE AND APPLIED MATHEMATICS
Vol. 3, No. 2, 2010, 141-162
ISSN 1307-5543 – www.ejpam.com

Stepwise Variable Selection for Loglinear Mixtures in Record Linkage

Rong Zhu [1,*], Jin Zhang [2], Da Zhang [3], Guohua Yan [4]

[1] *Department of Mathematics and Statistics, McMaster University, Canada*
[2] *Office of Institutional Research and Analysis, McMaster University, Canada*
[3] *Department of Computer Science Technology, Ohio University Lancaster, USA*
[4] *Department of Mathematics and Statistics, University of New Brunswick, Canada*

Abstract. A model building strategy is proposed to improve the probabilistic match in record linkage with focus on the loglinear mixture model of two components, each for the matched and unmatched pairs respectively. In reality, comparison attributes (i.e., covariates) often interact with each other, leading to more or less interactions in the loglinear models for both the matched and unmatched pairs. However, the interactions patterns are often not the same for both components. Particularly, because the number of matched pairs is usually very small compared with that of unmatched pairs in practice, the model for matched pairs can not be fitted with the same higher order interactions as that for the unmatched pairs. The proposed strategy is data-driven, and attempts to avoid both underfitting and overfitting due to subjective model specification for the data. Starting from the situation of no interaction, we add interactions sequentially in two loglinear components using the forward selection approach. Specifically, we define the alternatively climbing pathways through mixture families of two components with higher order interactions. The mixture models expanded along a pathway are nested successively. Thus, conventional tests used for comparison of nested models can be applied. Regarding parameter estimation for the mixture, a simplified method (including the choice of initial values of parameters) for the EM algorithm is developed, which facilitates the mixture model fitting using existing packages and functions in sophisticated statistical software like R. Simulation studies have then been conducted for various situations to assess the model selection approach, and comparisons of the selected models with the naive model assuming field independence have been made. We have applied this strategy to the record linkage case study in 2006 Annual Meeting of Statistical Society of Canada (SSC) and identified interactions among certain comparison attributes for both matched and unmatched pairs; these interactions are not always the same for both mixture components.

2000 Mathematics Subject Classifications: 62F07, 62F10, 62J99, 62P25

Key Words and Phrases: record linkage, loglinear mixture models, EM algorithm, model selection, alternatively climbing pathways

*Corresponding author.

Email addresses: rzhu@math.mcmaster.ca (R. Zhu), zhangjn@mcmaster.ca (J. Zhang), zhangd1@ohio.edu (D. Zhang), gyan@unb.ca (G. Yan)

1. Introduction

Information often comes from various sources regarding different aspects of objects, and could be recorded at different time points. For example, the provincial resident registry information is comprised of several data sources from different departments such as transportation, health, license and registration. These data sets include specific personal information such as name, date of birth and address. Typically, there is no unique ID assigned to individual objects among all data sources. However, for some reasons, like administration or detection of particular interested individuals (say, terrorists), the need arises to link two or more data sources together so that the records for the same individual object across different files can be identified for further processing. This task also identifies duplicate records in one file where a second virtual file is the copy of the original one.

The way to decide whether two records are the same or whether two records describe the same object is called record linkage. Records in different data files consist of common attributes (such as name, address and date of birth) and different attributes (such as weight in one file and height in another file) for individual objects. Usually, the common attributes are chosen to be identifiers for record matching. However, the different attributes in two files may provide partial information for the matching if these attributes are highly associated. In this study, we only consider the common attributes as the identifiers. Comparison vectors are then obtained for each record pair by comparing the values of identifiers.

The status of a pair of records for comparison is either "matched" or "unmatched". Ideally, if the values of identifiers match exactly for a record pair, then the two records are from the same object. However, in reality, there are many situations where the values of identifiers may not be the same even if they are from the same object. For example, the surname of a female may change when her marriage status changes. Hence, the method of exact match or deterministic match may miss many matched records. On the other hand, exact match may not always imply the same object. For instance, if the identifiers are just name and date of birth, it could occur that two different persons have the same name and date of birth although the chance is small. For this reason, probabilistic match is employed where the pair of records is regarded as from the same object if the weight involving the matched and unmatched probabilities is large enough.

Historically, Newcombe *et al*. (1959) studied probabilistic linkage on vital records; refer to [13]. Fellegi and Sunter (1969) developed the formal mathematical framework for probabilistic record linkage, and they also specified models in a hierarchical way according to a possible error-making mechanism; see [6]. After that, many techniques and software have been developed for record linkage. For example, blocking is used to reduce not only the size of comparison pairs, but also the dependence effects (i.e., interaction between identifiers). Nowadays record linkage methodology is used in many areas such as census, survey, administration and medical research; see [1, 3, 8, 12].

Regarding modelling, the approach applying a mixture of two components to the matched and unmatched record pairs seems to be more convenient, because it does not require the details of the error-making mechanism. Jaro (1989) considered a mixture model for the simplest contingency table derived from comparison vectors where fields are binary and independent,

and used the EM algorithm for parameter estimation with application in a test census of Tampa in 1985; see [7]. Winkler (1989) presented a loglinear model to adjust the lack of independence; refer to [16]. More work on loglinear models can be found in [8, 9, 10], and references therein.

We focus on mixture models. For a given record linkage task, many mixture models can be specified. However, which model fits the observed comparison pairs better, and how can underfitting or overfitting be avoided? These are questions for record linkage practitioners. Our study is motivated by the record linkage case study in 2006 Annual Meeting of Statistical Society of Canada (SSC 2006). We tried some free software for record linkage and found that their models assume field independence without interactions. The possible reason for this limitation could be the computational burden for arbitrary mixture models. However, the lack of fit is obvious in this simple model if fields are dependent. On the other hand, subjectively specifying models with higher order interactions may lead to overfitting for one or both components. Both underfitting and overfitting will affect the calculation of matched and unmatched probabilities, and thus affect the partition of record pairs. These problems lead to the model selection in the mixture setting.

In our study, we propose a strategy for model selection in loglinear mixtures, where parameter estimates are obtained via a simplified method for the EM algorithm for the mixture of two components. Such a simplified method can take advantage of existing packages and functions in sophisticated statistical software like R, so computationally it is tractable. We define the alternatively climbing pathways to indicate a sequence of nested models in the mixture family. For comparison, we adopt the Pearson Chi-square test for nested models on alternatively climbing pathways. The feature of this strategy is data-driven, and thus, the model building avoids subjective specification. Simulation studies have been conducted to verify the strategy, and to acquire experience of model building. We compare the results from this strategy with those from the simplest model where all factors are binary and independent. It shows that the data-driven strategy does give a refinement on modelling and lead to more accurate estimation of probabilities compared with the simplest model.

This paper is organized as follows. We outline statistical modelling and a simplified method for the EM algorithm in record linkage in Section 2. In Section 3, we propose the data-driven strategy for modelling loglinear mixtures. Then we conduct simulation studies to assess the proposed strategy in Section 4, and apply it to a case study in Section 5. Finally, we make brief concluding remarks in Section 6.

2. Statistical framework and parameter estimation for record linkage

In this section, we first outline the statistical framework for the record linkage, where the population of all pairs is the mixture of two components or subpopulations: matched pairs and unmatched pairs. Then we give a simplified method for the EM algorithm for this specific type of mixture models.

2.1. Notation and framework

Consider two files where file A has a records and file B has b records. The total number of possible record pairs for comparison is $N = a \times b$. The records consist of attributes which can be categorical variables such as name, address and date of birth, or numerical variables such as age. Select n common attributes as the identifiers for comparing pairs of records from both files A and B. Then the comparison for each record pair is made for all n identifiers by matching rules defined according to the needs for these attributes. This will result in a comparison vector of length n with each component being categorical. Each component associated with an identifier (attribute) in the comparison vector is called a field. The field values of comparison vector could be binary like 0 and 1, or any others defined by some rules. For example, the simplest rule is agreement/disagreement on an attribute. Under this simplest rule, if comparison attributes are name, address, gender and date of birth, an observed comparison vector $(0,1,1,0)$ means disagreement on name and date of birth, but agreement on address and gender. Thus, we obtain N comparison vectors for all record pairs denoted as follows:

$$\boldsymbol{\gamma}^j = (\gamma_1^j, \ldots, \gamma_n^j), \qquad j = 1, \ldots, N,$$

where the subscript denotes the comparison field while the superscript indicates the comparison pair.

Note that all comparison pairs include two types of pairs: matched and unmatched. Therefore, we can view the entire population of all possible comparison pairs as a mixture of two components or subpopulations: matched pairs and unmatched pairs. However, in this mixture, the membership for each pair is missing or unknown.

Let M denote the matched subpopulation and U denote the unmatched subpopulation. Each pair is either from M or U. Let

$$m_j = P(\boldsymbol{\gamma}^j | M), \qquad u_j = P(\boldsymbol{\gamma}^j | U), \qquad j = 1, \ldots, N.$$

Define the matching weight as follows:

$$w_j = \log \frac{P(\boldsymbol{\gamma}^j | M)}{P(\boldsymbol{\gamma}^j | U)} = \log P(\boldsymbol{\gamma}^j | M) - \log P(\boldsymbol{\gamma}^j | U), \quad j = 1, \ldots, N,$$

the logarithm of the likelihood ratio between the matched and unmatched. Probabilistic models are specified to obtain the associated probabilities.

Fellegi and Sunter (1969) set the fundamental probabilistic framework for record linkage; see [6]. Assume that m_j, u_j, w_j are estimated from data. Without loss of generality, the descending matching weight sequence is assumed to be

$$w_1 \geq \ldots \geq w_k \geq \ldots \geq w_l \geq \ldots \geq w_N,$$

and the corresponding matched and unmatched probability sequences are

$$m_1, \ldots, m_k, | \ \ldots, | \ m_l, \ldots, m_N,$$

$$u_1,\ldots,u_k,|\ \ \ldots,\ |\,u_l,\ldots,u_N,$$

where k and l are the cutting points which partition all comparison pairs into three parts: matched, uncertain and unmatched, as follows:

$$\begin{aligned} C_1:&\quad j=1,\ldots,k;\quad \text{matched},\\ C_2:&\quad j=k+1,\ldots,l-1;\quad \text{uncertain for clerical review},\\ C_3:&\quad j=l,\ldots,N;\quad \text{unmatched}. \end{aligned}$$

The cutting conditions are set as

$$\sum_{j=1}^{k} u_j \le \alpha_1,\quad \sum_{j=l}^{N} m_j \le \alpha_2,\quad \text{where } \alpha_1 \text{ and } \alpha_2 \text{ are prespecified.}$$

A good tutorial was given by Fair and Whitridge (1997); see [5]. In addition, Winkler (2005) gave an overview on record linkage; refer to [17].

In the literature, various models have been proposed for the mixture population from different angles. We have mentioned a few in Section 1. In this study, we focus on loglinear mixture models. The advantage of using a parametric model is that we can ignore the underlying error-making mechanism. Note that the mixture models for record linkage consist of only two components. In next subsection, we will outline the general framework for this particular type of mixture models and propose a simplified method for the EM algorithm as well as the choice of initial parameter values for this method.

2.2. Mixture models and the simplified method of the EM algorithm for parameter estimation

We consider parametric models for both the matched and unmatched subpopulations. Assume that the model for M has pmf or pdf $f_M(\boldsymbol{\gamma};\boldsymbol{\beta}_m)$ while the model for U has pmf or pdf $f_U(\boldsymbol{\gamma};\boldsymbol{\beta}_u)$. Here $\boldsymbol{\gamma}$ is the comparison vector, and $\boldsymbol{\beta}_m$ and $\boldsymbol{\beta}_u$ are parameter vectors associated with the models for M and U respectively. Suppose the chance of a comparison pair from M is π. Then the pmf or pdf for the j-th comparison pair is

$$f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}^j;\boldsymbol{\beta}_m,\boldsymbol{\beta}_u,\pi)=\pi f_M(\boldsymbol{\gamma}^j;\boldsymbol{\beta}_m)+(1-\pi)f_U(\boldsymbol{\gamma}^j;\boldsymbol{\beta}_u). \tag{1}$$

On the other hand, denote the membership for the j-th comparison pair as follows

$$g_j=\begin{cases}1, & \text{if } \boldsymbol{\gamma}^j \text{ is from the matched population } M,\\ 0, & \text{if } \boldsymbol{\gamma}^j \text{ is from the unmatched } U,\end{cases}\qquad j=1,\ldots,N.$$

For the j-th pair, the complete data is $\mathbf{X}_j=(\boldsymbol{\gamma}^j,g_j)$. Thus, the pmf or pdf for the complete data of the j-th pair is

$$f_{\mathbf{X}}\left((\boldsymbol{\gamma}^j,g_j);\boldsymbol{\beta}_m,\boldsymbol{\beta}_u,\pi\right)=\left[\pi f_M(\boldsymbol{\gamma}^j;\boldsymbol{\beta}_m)\right]^{g_j}\left[(1-\pi)f_U(\boldsymbol{\gamma}^j;\boldsymbol{\beta}_u)\right]^{(1-g_j)}. \tag{2}$$

Assume that the comparison pairs are independent. The logarithm of the joint pmf or pdf for all complete data is

$$\begin{aligned} l_C\left((\boldsymbol{\gamma}^1, g_1), \ldots, (\boldsymbol{\gamma}^N, g_N); \boldsymbol{\beta}_m, \boldsymbol{\beta}_u, \pi\right) &= \log\left(\prod_{j=1}^{N} f_{\mathbf{X}}\left((\boldsymbol{\gamma}^j, g_j); \boldsymbol{\beta}_m, \boldsymbol{\beta}_u, \pi\right)\right) \\ &= \sum_{j=1}^{N}\Big[g_j \log f_M(\boldsymbol{\gamma}^j; \boldsymbol{\beta}_m) + (1-g_j)\log f_U(\boldsymbol{\gamma}^j; \boldsymbol{\beta}_u) \\ &\qquad + g_j \log \pi + (1-g_j)\log(1-\pi)\Big]. \end{aligned} \tag{3}$$

The conditional pmf for the membership g is then

$$\begin{aligned} f_{g|\gamma}\left(g_j \mid \boldsymbol{\gamma}^j; \boldsymbol{\beta}_m, \boldsymbol{\beta}_u, \pi\right) &= \frac{f_{\mathbf{X}}\left((\boldsymbol{\gamma}^j, g_j); \boldsymbol{\beta}_m, \boldsymbol{\beta}_u, \pi\right)}{f_{\gamma}(\boldsymbol{\gamma}^j; \boldsymbol{\beta}_m, \boldsymbol{\beta}_u, \pi)} \\ &= \frac{f_{\mathbf{X}}\left((\boldsymbol{\gamma}^j, g_j); \boldsymbol{\beta}_m, \boldsymbol{\beta}_u, \pi\right)}{\pi f_M(\boldsymbol{\gamma}^j; \boldsymbol{\beta}_m) + (1-\pi) f_U(\boldsymbol{\gamma}^j; \boldsymbol{\beta}_u)}. \end{aligned} \tag{4}$$

However, the membership g_j is missing in the mixture population. Therefore, the mixture can be regarded as a missing value problem and the EM algorithm can be utilized to estimate the model parameters $\boldsymbol{\beta}_m$, $\boldsymbol{\beta}_u$ and π. For references on the EM algorithm, see [4], [11] and references therein.

The EM algorithm is an iterative approach which alternates between two steps, an E-step and an M-step. The E-step is conducted as follows. Conditional on values of $\boldsymbol{\beta}_m^{(k)}$, $\boldsymbol{\beta}_u^{(k)}$ and $\pi^{(k)}$, as well as the observed comparison vectors, the expectation of the logarithm of the joint pmf or pdf for all complete data is

$$\begin{aligned} L &= \mathbf{E}\left[l_C\left((\boldsymbol{\gamma}^1, g_1), \ldots, (\boldsymbol{\gamma}^N, g_N); \boldsymbol{\beta}_m, \boldsymbol{\beta}_u, \pi\right) \mid \boldsymbol{\gamma}^1, \ldots, \boldsymbol{\gamma}^N; \boldsymbol{\beta}_m^{(k)}, \boldsymbol{\beta}_u^{(k)}, \pi^{(k)}\right] \\ &= \sum_{j=1}^{N} \mathbf{E}\left[\log\left(f_{\mathbf{X}}\left((\boldsymbol{\gamma}^j, g_j); \boldsymbol{\beta}_m, \boldsymbol{\beta}_u, \pi\right)\right) \mid \boldsymbol{\gamma}^j; \boldsymbol{\beta}_m^{(k)}, \boldsymbol{\beta}_u^{(k)}, \pi^{(k)}\right] \\ &= \sum_{j=1}^{N} \int \log\left[f_{\mathbf{X}}\left((\boldsymbol{\gamma}^j, g_j); \boldsymbol{\beta}_m, \boldsymbol{\beta}_u, \pi\right)\right] f_{g|\gamma}\left(g_j \mid \boldsymbol{\gamma}^j; \boldsymbol{\beta}_m^{(k)}, \boldsymbol{\beta}_u^{(k)}, \pi^{(k)}\right) dg_j \\ &= \sum_{j=1}^{N} \sum_{g_j=0}^{1} \Big\{\Big[g_j \log f_M(\boldsymbol{\gamma}^j; \boldsymbol{\beta}_m) + (1-g_j)\log f_U(\boldsymbol{\gamma}^j; \boldsymbol{\beta}_u) \\ &\qquad + g_j \log \pi + (1-g_j)\log(1-\pi)\Big] \times f_{g|\gamma}\left(g_j \mid \boldsymbol{\gamma}^j; \boldsymbol{\beta}_m^{(k)}, \boldsymbol{\beta}_u^{(k)}, \pi^{(k)}\right)\Big\}. \end{aligned}$$

Let

$$g_m^{(k)}(\boldsymbol{\gamma}^j) = f_{g|\gamma}\left(1 \mid \boldsymbol{\gamma}^j; \boldsymbol{\beta}_m^{(k)}, \boldsymbol{\beta}_u^{(k)}, \pi^{(k)}\right) = \frac{\pi^{(k)} f_M(\boldsymbol{\gamma}^j; \boldsymbol{\beta}_m^{(k)})}{\pi^{(k)} f_M(\boldsymbol{\gamma}^j; \boldsymbol{\beta}_m^{(k)}) + (1-\pi^{(k)}) f_U(\boldsymbol{\gamma}^j; \boldsymbol{\beta}_u^{(k)})},$$

$$\begin{aligned} g_u^{(k)}(\boldsymbol{\gamma}^j) &= f_{g|\gamma}\left(0 \mid \boldsymbol{\gamma}^j; \boldsymbol{\beta}_m^{(k)}, \boldsymbol{\beta}_u^{(k)}, \pi^{(k)}\right) = \frac{(1-\pi^{(k)}) f_U(\boldsymbol{\gamma}^j; \boldsymbol{\beta}_u^{(k)})}{\pi^{(k)} f_M(\boldsymbol{\gamma}^j; \boldsymbol{\beta}_m^{(k)}) + (1-\pi^{(k)}) f_U(\boldsymbol{\gamma}^j; \boldsymbol{\beta}_u^{(k)})} \\ &= 1 - g_m(\boldsymbol{\gamma}^j). \end{aligned}$$

Then, by algebra,

$$\begin{aligned} L &= \sum_{j=1}^{N} g_m^{(k)}(\boldsymbol{\gamma}^j) \log f_M(\boldsymbol{\gamma}^j; \boldsymbol{\beta}_m) + \sum_{j=1}^{N} g_u^{(k)}(\boldsymbol{\gamma}^j) \log f_U(\boldsymbol{\gamma}^j; \boldsymbol{\beta}_u) \\ &\qquad + \sum_{j=1}^{N} \left[g_m^{(k)}(\boldsymbol{\gamma}^j) \log \pi + g_u^{(k)}(\boldsymbol{\gamma}^j) \log(1-\pi) \right] \\ &= L_1 + L_2 + L_3, \end{aligned}$$

where

$$L_1 = \sum_{j=1}^{N} g_m^{(k)}(\boldsymbol{\gamma}^j) \log f_M(\boldsymbol{\gamma}^j; \boldsymbol{\beta}_m), \qquad L_2 = \sum_{j=1}^{N} g_u^{(k)}(\boldsymbol{\gamma}^j) \log f_U(\boldsymbol{\gamma}^j; \boldsymbol{\beta}_u),$$

and

$$L_3 = \sum_{j=1}^{N} \left[g_m^{(k)}(\boldsymbol{\gamma}^j) \log \pi + g_u^{(k)}(\boldsymbol{\gamma}^j) \log(1-\pi) \right].$$

This splits the conditional expectation L into three separate parts corresponding to the matched pairs, the unmatched pairs and the mixing proportion π respectively.

The M-step maximizes the conditional expectation L, which is in turn equivalent to maximization of L_1, L_2 and L_3 separately. Note that L_1 and L_2 are weighted log-likelihoods for the matched and unmatched pairs respectively. Thus, for models which can be fitted by weighted MLE using available software, their maximization can be readily obtained by that software. The maximization of L_3 is straightforward:

$$\frac{\partial L_3}{\partial \pi} = \sum_{j=1}^{N} \left[g_m^{(k)}(\boldsymbol{\gamma}^j) \frac{1}{\pi} - g_u^{(k)}(\boldsymbol{\gamma}^j) \frac{1}{1-\pi} \right] = 0$$

yields

$$\frac{1}{\pi} \sum_{j=1}^{N} g_m^{(k)}(\boldsymbol{\gamma}^j) = \frac{1}{1-\pi} \sum_{j=1}^{N} g_u^{(k)}(\boldsymbol{\gamma}^j).$$

Thus,

$$\pi = \frac{\sum_{j=1}^{N} g_m^{(k)}(\boldsymbol{\gamma}^j)}{\sum_{j=1}^{N} g_m^{(k)}(\boldsymbol{\gamma}^j) + \sum_{j=1}^{N} g_u^{(k)}(\boldsymbol{\gamma}^j)} = \frac{1}{N} \sum_{j=1}^{N} g_m^{(k)}(\boldsymbol{\gamma}^j).$$

We obtain

$$\pi^{(k+1)} = \frac{1}{N}\sum_{j=1}^{N} g_m^{(k)}(\boldsymbol{\gamma}^j), \quad \boldsymbol{\beta}_m^{(k+1)} = \arg\max L_1, \quad \boldsymbol{\beta}_u^{(k+1)} = \arg\max L_2. \tag{5}$$

This approach is regarded as a simplified method of the EM algorithm for the mixture situation. It can take advantage of existing packages or functions in statistical software without developing particular code for maximization of the complicated model components, especially with interactions. Note that Winkler (1988) investigated the EM algorithm for Fellegi-Sunter model; see [15]. We are suggestting a computational method for fitting a mixture with two components where each component is readily fitted by a model like GLM.

3. Model building strategy for loglinear mixtures of two components

We consider all fields of the comparison vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$ are categorical. For example, γ_1 could be binary taking value 0 or 1, meaning unmatching or matching respectively for a pair of records in field 1. Then all of such observed comparison vectors form a multi-way contingency table with n factors (each field corresponds to a factor). Thus, it is natural to utilize the loglinear model for analyzing such a contingency table.

A loglinear model states that the logarithm of the expected number of a cell in the contingency table can be expressed as the additive function of main effects and interactions of factors. For instance, we consider a four factor $I \times J \times K \times L$ contingency table. Two loglinear models we will use in the remaining are

$$\log \mu_{ijkl} = \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_l^{(4)}, \tag{6}$$

or

$$\log \mu_{ijkl} = \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_l^{(4)} + \lambda_{ij}^{(12)} + \lambda_{ik}^{(13)} + \lambda_{il}^{(14)} + \lambda_{jk}^{(23)} + \lambda_{jl}^{(24)} + \lambda_{kl}^{(34)}, \tag{7}$$

$$i = 1, \ldots, I; \quad j = 1, \ldots, J; \quad k = 1, \ldots, K, \quad l = 1, \ldots, L.$$

Here μ_{ijkl} denotes the expected cell count, λ is the overall mean, $\lambda_i^{(1)}$, $\lambda_j^{(2)}$, $\lambda_k^{(3)}$ and $\lambda_l^{(4)}$ are the main effects of each factor at corresponding levels i, j, k and l, $\lambda_{ij}^{(12)}$, $\lambda_{ik}^{(13)}$, $\lambda_{il}^{(14)}$, $\lambda_{jk}^{(23)}$, $\lambda_{jl}^{(24)}$, $\lambda_{kl}^{(34)}$ are two-factor interactions at specified level pairs. Model (6) corresponds to the situation of independence among factors, while model (7) corresponds to a situation of dependence among factors. Note that model (6) is nested in model (7).

Apart from the above common ANOVA-like formulation, there is a generalized linear model formulation, where the cell counts are assumed to be distributed in Poisson. For instance, Model (6) and (7) have the following equivalent Poisson GLM formulations:

$$\log \mu = \alpha + \alpha_1\gamma_1 + \alpha_2\gamma_2 + \alpha_3\gamma_3 + \alpha_4\gamma_4, \tag{8}$$

and

$$\log \mu \quad = \quad \alpha + \alpha_1\gamma_1 + \alpha_2\gamma_2 + \alpha_3\gamma_3 + \alpha_4\gamma_4$$

$$+\alpha_{12}\gamma_1\gamma_2+\alpha_{13}\gamma_1\gamma_3+\alpha_{14}\gamma_1\gamma_4+\alpha_{23}\gamma_2\gamma_3+\alpha_{24}\gamma_2\gamma_4+\alpha_{34}\gamma_3\gamma_4. \quad (9)$$

Here α's are the intercept for the model and coefficients for variable γ's and their interactions, μ is the expected cell count associated with covariates $\boldsymbol{\gamma}=(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$.

R has function `loglm()` for fitting loglinear models in the form of ANOVA-like specification, and function `glm()` for fitting Poisson GLM models. Each has advantages and disadvantages. We use both in our computation in case one encounters unexpected problems.

In record linkage, the matched and unmatched subpopulations could have their own model specifications, thus, M and U are assumed to have different loglinear models respectively. However, the membership is unknown. So the entire population of possible pairs is a mixture of two loglinear models for M and U. The observed contingency table is a sample from this mixture population.

Jaro (1989) considered the simplest case for the contingency table where all factors are binary, and thus, the contingency table is $2\times 2\times\cdots\times 2$, and the number of all cells of the table is 2^n; see [7]. He assumed field independence, and employed the EM algorithm for the estimation of all probabilities related to fields and the mixing proportion. From the prospective of loglinear model formulation, the models for M and U in Jaro (1989) were loglinear models without interactions. The field independence may not hold in reality. For instance, strong dependencies were observed by Thibaudeau (1993) in census data; see [14]. Prior to that report, Winkler (1989) extended the simplest case to a general one, where interactions were taken into account in the models for M and U, specifically, two-class and three-class interactions were considered, and models were assessed using goodness-of-fit tests (see [16]). Larsen (1997) considered a loglinear model with main effects, all two-way interactions, all three-way interactions and two five-way interactions in a trial census and post-enumeration survey (see [9]).

Any of the specified loglinear mixture models can be estimated using the simplified method for the EM algorithm in Section 2. For a real case, the specification can be given from background information or subjective opinion. Thus, different opinions or understanding about the real case can result in different model specifications.

However, among those subjectively specified models, which one fits the observed contingency table better, or captures the data feature better? How can we avoid overfitting or underfitting? These questions lead to the model selection for loglinear mixture models in record linkage. Usually, a better model will give a better estimation of cell probabilities for matched and unmatched pairs, which in turn increases the accuracy of record pair partition.

When we search among a series of models, a method of statistical testing is required for model comparison at each step, and a strategy is needed in the iterative selection process. Suppose the loglinear mixture model $\mathscr{M}^{(0)}$ is nested in another loglinear mixture model $\mathscr{M}^{(1)}$, namely $\mathscr{M}^{(0)}\subset\mathscr{M}^{(1)}$. We adopt a general method to compare these two loglinear mixture models, i.e., testing

$$H_0:\ \mathscr{M}^{(0)}=\mathscr{M}^{(1)} \quad \text{versus} \quad H_1:\ \mathscr{M}^{(0)}\subset\mathscr{M}^{(1)}.$$

If H_0 is not rejected, we do not need to improve $\mathscr{M}^{(0)}$ by considering the larger model $\mathscr{M}^{(1)}$. Otherwise, model refinement is necessary. For this testing, the Pearson Chi-squared statistic is

employed as follows

$$\chi^2\left(\mathscr{M}^{(0)}|\mathscr{M}^{(1)}\right)=\sum(\hat{\mu}_{1i}-\hat{\mu}_{0i})^2/\hat{\mu}_{0i}, \tag{10}$$

where $\hat{\mu}_{1i}$ and $\hat{\mu}_{0i}$ are fitted values of individual cell expectations for $\mathscr{M}^{(0)}$ and $\mathscr{M}^{(1)}$ respectively. Because one model is nested in the other, the asymptotic distribution of $\chi^2\left(\mathscr{M}^{(0)}|\mathscr{M}^{(1)}\right)$ under H_0 is $\chi^2(df)$, where the degrees of freedom df is the difference of parameter numbers between $\mathscr{M}^{(0)}$ and $\mathscr{M}^{(1)}$. For a reference, see [2], P. 364.

The loglinear models for components U and M in the mixture we consider for record linkage may or may not include higher order interactions among fields. We denote $\mathscr{M}_{ij}$ for a loglinear mixture model where i indicates that the loglinear model for U has up to the $(i+1)$-th order interactions (i.e., $(i+1)$-factor-interactions) and j indicates that the loglinear model for M has up to the $(j+1)$-th order interactions (i.e., $(j+1)$-factor-interactions), $i,j\geq 0$. For instance, $\mathscr{M}_{00}$ means that the models for both M and U in the mixture have no interactions, i.e., Model (6). $\mathscr{M}_{10}$ means that the model for U has two-factor interactions, while the model for M has no interactions. $\mathscr{M}_{02}$ means that the model for U has no interactions, while the model for M has two-factor and three-factor interactions. The nested situation of two loglinear mixture models is totally determined by the nested situations of their both components. Therefore, we can build nested loglinear mixture models by building nested loglinear models for either U or M, or both U and M. That is,

$$\mathscr{M}_{ij}\subset\mathscr{M}_{i'j'}\quad \text{if } i\leq i' \text{ and } j\leq j'.$$

For instance, $\mathscr{M}_{00}\subset\mathscr{M}_{01}\subset\mathscr{M}_{11}$ and $\mathscr{M}_{00}\subset\mathscr{M}_{10}\subset\mathscr{M}_{11}$, but $\mathscr{M}_{01}$ and $\mathscr{M}_{10}$ are not nested in each other. See the illustration in (11) where mixture models of two components are arranged in a matrix format.

$$\begin{array}{ccccccc}
\vdots & & \vdots & & \vdots & & \vdots \\
\mathscr{M}_{30} & & \mathscr{M}_{31} & & \mathscr{M}_{32} & \rightarrow & \mathscr{M}_{33} \quad \cdots \\
 & & & & \uparrow & & \Uparrow \\
\mathscr{M}_{20} & & \mathscr{M}_{21} & \rightarrow & \mathscr{M}_{22} & \Rightarrow & \mathscr{M}_{23} \quad \cdots \\
 & & \uparrow & & \Uparrow & & \\
\mathscr{M}_{10} & \rightarrow & \mathscr{M}_{11} & \Rightarrow & \mathscr{M}_{12} & & \mathscr{M}_{13} \quad \cdots \\
\uparrow & & \Uparrow & & & & \\
\mathscr{M}_{00} & \Rightarrow & \mathscr{M}_{01} & & \mathscr{M}_{02} & & \mathscr{M}_{03} \quad \cdots
\end{array} \tag{11}$$

We make comparison for models on the nested pathways. However, there are many nested paths in the family of loglinear mixtures of two components. For example, from $\mathscr{M}_{00}$ to $\mathscr{M}_{11}$, there are two nested pathways: $\mathscr{M}_{00}\subset\mathscr{M}_{10}\subset\mathscr{M}_{11}$ and $\mathscr{M}_{00}\subset\mathscr{M}_{01}\subset\mathscr{M}_{11}$. In general, from one diagonal model $\mathscr{M}_{ii}$ to the next diagonal model $\mathscr{M}_{(i+1)(i+1)}$, there are two nested pathways similar to this illustrated example. Refer to (11). This pair of nested pathways is called alternatively climbing pathways (denoted by "$\rightarrow$" and "$\Rightarrow$" respectively). We propose the following forward selection strategy along the alternatively climbing pathways:

- Step 1: Start from the loglinear mixture model $\mathscr{M}_{00}$ (the one on the bottom left corner in (11)), where the loglinear models for both U and M have only main factor effects with no interactions.

- Step 2: Proceed on the pathway $\mathscr{M}_{00} \subset \mathscr{M}_{10} \subset \mathscr{M}_{11}$, and compare two adjacent models successively using a prespecified significance level α. This pathway considers the model with higher order interactions for U first.
  - If the series testing stops at $\mathscr{M}_{00}$, then go to step 3.
  - If the series testing stops at $\mathscr{M}_{10}$, then the final model is $\mathscr{M}_{10}$.
  - If the series testing stops at $\mathscr{M}_{11}$, then go to step 4.
- Step 3: Proceed on the alternative pathway $\mathscr{M}_{00} \subset \mathscr{M}_{01} \subset \mathscr{M}_{11}$, and compare two adjacent models successively using the significance level α. This pathway considers the model with higher order interactions for M first.
  - If the series testing stops at $\mathscr{M}_{00}$, then the final model is $\mathscr{M}_{00}$.
  - If the series testing stops at $\mathscr{M}_{01}$, then the final model is $\mathscr{M}_{01}$.
  - If the series testing stops at $\mathscr{M}_{11}$, then go to step 4.
- Step 4: Treat $\mathscr{M}_{11}$ like $\mathscr{M}_{00}$, and repeat steps 1 to 3 along the alternatively climbing pathways from $\mathscr{M}_{11}$ to $\mathscr{M}_{22}$, i.e., start a second round of search from the next diagonal entry.
- Step 5: Continue the iteration until the series testing stops.

The above search on alternatively climbing pathways guarantees the nesting between two loglinear mixture models under comparison. We place priority on the pathway where the model for U is added higher order interactions (the route denoted by "$\rightarrow$") first, because in record linkage, the unmatched pairs are the majority. If this fails, then we consider the other pathway where the model for M is added higher order interactions (the route denoted by "$\Rightarrow$").

Note that for step 1, if the contingency table degenerates to the case where all factors are binary, the computation will become the Jaro's method. The simplest loglinear mixture model $\mathscr{M}_{00}$ captures the main effects explained by factors alone. Note that the attributes chosen by investigators are typically important variables in record matching according to background knowledge. Thus, their main effects practically exist. On the other hand, we may not know the joint effects of attributes in advance. Thus, it is safe and convenient to start from the simplest mixture model $\mathscr{M}_{00}$.

However, if the joint effects have been known to be strong by the background knowledge, we can start from a particular mixture model which may not be a diagonal entry in (11). A modification of the model search is then that we proceed with the testings on an upward or a rightward pathway to the model in the closest diagonal entry, then continue the search using the above strategy. Note that the backward selection could be feasible. However, it starts from the most complicated model with all possible interactions, and thus, the computational cost is typically large.

In record linkage, the number of matched pairs is extremely small relative to the number of unmatched pairs. So usually the model for M has lower order interactions than that for U. That is why we consider to enlarge the component for U first. The restriction of searching

along the alternatively climbing pathways prevents over-extension of model for either component U or M. This conservative procedure is based on the reasoning that interactions among attributes usually affect both subpopulations U and M. Thus, we do not wish the model for U has higher order interactions while the model for M does not. If this reasoning is not appropriate, the restriction can be relaxed, say we can further extend the model for U when a final model is $M_{(i+1)i}$. However, our simulation studies show that it is hard to capture higher order interactions from data even if they are truly generated from a model with higher order interactions. Therefore, from the practical viewpoint, we suggest not to seek very higher order interactions in record linkage.

The final model obtained from the proposed strategy will include all interactions of the same order, say all three-factor-interactions or third order interactions. It may not be necessary to do so because some of them could be very weak. Thus, we can apply testing in (10) again to prune unnecessary interactions for either U or M in the final mixture model.

The choice of initial values of parameters for the EM algorithm is important for the final solution. If the contingency table is the one with binary factors, then we can use the Jaro's method to determine the initial values. Note that the maximum number of matched pairs is $\min(a, b)$. In a general case, we propose the following data-driven method named as the J method for the choice of initial values regarding parameters π, $\boldsymbol{\beta}_u$ and $\boldsymbol{\beta}_m$:

- the initial value for the mixing proportion is chosen as $\pi^{(0)} = \min(a, b)/ab = 1/\max(a, b)$, the upper bound of the mixing proportion. This can be modified as $\pi^{(0)}/10$, $\pi^{(0)}/5$ or $\pi^{(0)}/2$ depending on the blocking situation.
- $\boldsymbol{\beta}_u^{(0)}$, the vector of initial values in the loglinear model for U, is obtained from fitting that model to a modified table which has more unmatched pairs and fewer matched pairs. The cell values of this table are the original cell counts divided by $\min(a, b)$ (thus may not be integers).
- $\boldsymbol{\beta}_m^{(0)}$, the vector of initial values in the loglinear model for M, is obtained from fitting that model to another modified table which is created by picking counts for cells with higher matching chance, and assigning zeroes for others cells.

Intuitively, if the table is derived from more unmatched pairs or from more matched pairs, then the fitted model is closer to the true model for U or for M respectively. It works quite well in the simulation studies. Note that fitting a loglinear model to a non-integer-valued table is the same as fitting to an integer-valued table in R.

4. Simulation Studies

Simulation studies are necessary for verifying the proposed data-driven strategy of model building. From simulation, we can obtain experience on model fitting and building in various settings such as file sizes, mixing proportions and interactions patterns. Besides, we can also identify potential drawbacks for necessary adjustment.

We consider the comparison vector consisting of four fields, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3, \gamma_4)$, which is fairly moderate in reality. Each field is binary, where 0 means the field unmatched while 1

indicates the field matched. Thus, the total number of cells is $2^4 = 16$. Table 1 labels all of these sixteen cells in the way of a 2^4 factorial design matrix for later reference. Commonly,

Table 1: Cell labels for the $2 \times 2 \times 2 \times 2$ contingency table.

Cell	Comparison vector	Cell	Comparison vector
1	$(0,0,0,0)$	9	$(0,0,0,1)$
2	$(1,0,0,0)$	10	$(1,0,0,1)$
3	$(0,1,0,0)$	11	$(0,1,0,1)$
4	$(1,1,0,0)$	12	$(1,1,0,1)$
5	$(0,0,1,0)$	13	$(0,0,1,1)$
6	$(1,0,1,0)$	14	$(1,0,1,1)$
7	$(0,1,1,0)$	15	$(0,1,1,1)$
8	$(1,1,1,0)$	16	$(1,1,1,1)$

a cell with more 1's has larger matching chance for record pairs with comparison vectors in that cell. Thus, cells corresponding to comparison vectors being $(1,1,1,1),(1,1,0,1)$ and $(1,1,1,0)$ have more matching possibilities than cells corresponding to $(0,0,0,0)$, $(0,0,1,0)$ and $(0,0,0,1)$.

This is the type of contingency table considered in Jaro (1989); see [7]. For this situation, a byproduct is that we can compare the estimated 16 cell probabilities from the model selected via the data-driven strategy with those from the simplest starting model $\mathscr{M}_{00}$. For each individual simulation, we set the lengths of two data sets as a and b, as well as the true matched pairs N_m (thus, leading to the mixing proportion π). For the sake of simplicity, we do not consider the blocking technique, thus, there are $N = ab$ record pairs. Six settings are then chosen for the simulation (see Table 2). The combination of lengths a and b covers three

Table 2: Lengths of two data sets and true matches in six setting cases.

Setting	1	2	3	4	5	6
a	200	200	400	400	1,000	1,000
b	1,000	1,000	400	400	2,000	2,000
$N = ab$	200,000	200,000	160,000	160,000	2,000,000	2,000,000
N_m	20	150	40	200	50	600

situations:

- one is small, and one is large;
- both are of equal lengths;
- both are large, but of different lengths.

Two situations regarding the mixing proportion are considered: one with a small mixing proportion corresponding to fewer matched pairs, and one with a fairly large proportion corresponding to more matched pairs. Regarding the loglinear mixture model, we choose two types of loglinear models for both U and M as follows.

(1) Models with no interactions for both U and M, i.e., Model $\mathscr{M}_{00}$. Specifications of main effects are in Table 3.

Table 3: Specifications of main effects of models with no interactions.

Component	λ	$\lambda_0^{(1)}$	$\lambda_1^{(1)}$	$\lambda_0^{(2)}$	$\lambda_1^{(2)}$	$\lambda_0^{(3)}$	$\lambda_1^{(3)}$	$\lambda_0^{(4)}$	$\lambda_1^{(4)}$
U	0.5	2.30	-2.30	1.59	-1.59	1.74	-1.74	2.09	-2.09
M	0.3	-1.10	1.10	-0.87	0.87	-0.69	0.69	-1.10	1.10

(2) Models with two-factor interactions for both U and M, i.e., Model $\mathscr{M}_{11}$. Specifications of main effects and interactions are in Table 4. Note that

$$\lambda_1^{(i)} = -\lambda_0^{(i)}, \quad \lambda_{01}^{(ij)} = \lambda_{01}^{(ij)} = -\lambda_{00}^{(ij)}, \quad \lambda_{11}^{(ij)} = \lambda_{00}^{(ij)}, \quad i,j = 1,2,3,4.$$

Thus, effects and interactions at other levels can be readily derived from the one at level 0.

Table 4: Specifications of main effects and interactions of models with interactions at level 0.

Component	λ	$\lambda_0^{(1)}$	$\lambda_0^{(2)}$	$\lambda_0^{(3)}$	$\lambda_0^{(4)}$	$\lambda_{00}^{(12)}$	$\lambda_{00}^{(13)}$	$\lambda_{00}^{(14)}$	$\lambda_{00}^{(23)}$	$\lambda_{00}^{(24)}$	$\lambda_{00}^{(34)}$
U	0.5	2.3	0.6	1.00	3.0	1.04	0.40	1.13	0.65	0.20	1.40
M	0.3	-1.2	-0.9	-0.66	-0.5	0.44	-0.64	-0.16	-1.08	1.08	0.32

With two different mixture models assigned for each of the six cases of data settings, we have investigated 12 simulation cases in total. Since the mixture model is for comparison vectors, we do not need to generate two original record files in each simulation case. Actually, what we need to generate are $(N - N_m)$ comparison vectors from the loglinear model for U and N_m comparison vectors from the loglinear model for M in each simulation case. These comparison vectors yield the $2 \times 2 \times 2 \times 2$ tables of 16 cells (see Table 1). Therefore, we finally need the contingency tables for U and M respectively. Note that conditional on the total count, the joint counts of all cells follow a multinomial distribution. Suppose Y_i and μ_i are the observed count and expected count in cell i respectively. Let N_0 be the total count. Then, conditional on $\sum_{i=1}^{16} Y_i = N_0$,

$$(Y_1, Y_2, \ldots, Y_{16}) \sim \text{Multinomial}(N_0; p_1, p_2, \ldots, p_{16}), \tag{12}$$

where $p_i = \mu_i / \sum_{j=1}^{16} \mu_j$, the probability that an observed vector occurs in cell i, $i = 1, 2, \ldots, 16$. For a reference, see [2], P. 317-318. In general, we can simulate a contingency table according to (12) for a specified total count. This is the method we use to simulate data sets for six settings with mixture model $\mathscr{M}_{11}$. However, if fields are independent, we can generate four independent binary columns of length being the specified total count N_0. Each column is a series of N_0 Bernoulli trials with success probability which can be readily computed according to (12). Information of a and b will be used in the choice of initial values in the simplified method for the EM algorithm.

We apply the proposed building strategy to select a loglinear mixture model for each simulation case. The significance level is set as $\alpha = 5\%$. The following lists the model building processes and the finally selected models for all the 12 simulation cases. Due to space limitation, we omit the details of the estimated models.

- For the true loglinear mixture model being $\mathscr{M}_{00}$ (refer to Table 3), settings 1, 2, 3, 5 and 6 lead to $\mathscr{M}_{00}$ while setting 4 yields $\mathscr{M}_{00} \to \mathscr{M}_{10}$. Thus, from the viewpoint of model fitting, setting 4 is a little bit overfitted with false interactions for U.
- For the true loglinear mixture model being $\mathscr{M}_{11}$ (refer to Table 4), settings 1, 2, 4 and 6 proceed as $\mathscr{M}_{00} \to \mathscr{M}_{10} \to \mathscr{M}_{11}$ while settings 3 and 5 result in $\mathscr{M}_{00} \to \mathscr{M}_{10}$. Hence, settings 3 and 5 are a little bit underfitted without capturing the two-factor interactions for M.

Regarding the true mixture model being $\mathscr{M}_{00}$, five out of six simulation cases yield the selected models consistent with the true one, however, one case results in overfitting. Further investigation shows that the p-values in testing $H_0 : \mathscr{M}_{00} = \mathscr{M}_{10}$ for the setting $a = 400$ and $b = 400$ are 0.047 (< 0.05) when $N_m = 40$ (setting 3) and 0.105 (> 0.05) when $N_m = 200$ (setting 4). Note that setting 4 has the highest proportion of matched pairs, with half of records in each file corresponding to the same objects. This highly matching feature might cause model inflation for component U when fields are actually independent. As to the true mixture model being $\mathscr{M}_{11}$, four out of six cases lead to models consistent with the true ones, but two cases fail to capture the two-factor interactions in the loglinear model for M, leading to underfitting for the mixture. Note that both overfitting and underfitting happen when the number of matched pairs N_m is small. Hence, empirically, it may be more difficult to find higher order interactions for the M component than for the U component in the mixture. Curiously, we have not observed any overfitting when the true mixture model is $\mathscr{M}_{11}$ in a larger scale investigation. These observed phenomena indicate that the proposed model building strategy may not capture the true underlying mixture model perfectly. Sometimes it yields a slight overfitting or underfitting for one of the mixture components.

Since in some cases the selected models deviate from the true models, a natural question is how the corresponding cell probabilities deviate from the truth. This is because the ultimate concern in record linkage is the probabilities from U and M for all record pairs, as well as their matching weights. Thus, it is necessary to investigate the cell probability differences. In addition, we can compare the selected model with the fitted starting model $\mathscr{M}_{00}$ from the

viewpoint of cell probabilities. This is to see whether or not our proposed data-driven strategy can beat or at least is equivalent to the subjective specification with naive field independence used in some record linkage software. We show the details in Table 5 for setting 5 when the true mixture model is $\mathscr{M}_{11}$. In this simulation case, the selected model by the data-driven strategy is underfitting in the M component, not perfectly consistent with the true underlying model. In Table 5,

- $PD(S,T|U)$ and $PD(S,T|M)$ denote differeences in cell probabilities between the selected model by the data-driven strategy and the true model for component U and M respectively,
- $PD(I,T|U)$ and $PD(I,T|M)$ denote differeences in cell probabilities between the simplest starting model $\mathscr{M}_{00}$ and the true model for component U and M respectively.

Table 5: Comparisons of cell probabilities between two fitted models and the true model, $\mathscr{M}_{11}$, for setting 5.

Cell	$PD(S,T\|U)$	$PD(I,T\|U)$	$PD(S,T\|M)$	$PD(I,T\|M)$
1	-2.710e-05	2.130e-04	-2.233e-04	4.487e-02
2	-2.547e-07	-3.767e-06	-5.058e-03	6.675e-03
3	2.973e-05	-8.666e-05	-5.604e-04	4.272e-01
4	2.606e-06	-2.535e-05	-7.308e-02	3.753e-02
5	2.865e-06	-1.356e-06	-1.375e-02	9.176e-03
6	5.579e-08	-2.366e-07	-2.407e-02	1.810e-02
7	-4.573e-06	-8.575e-05	-4.588e-04	2.170e-01
8	4.212e-07	-1.728e-06	-4.552e-03	5.192e-02
9	-2.060e-06	-3.412e-06	-5.084e-05	5.027e-03
10	-5.656e-08	-5.617e-08	-6.071e-04	7.143e-04
11	-5.870e-08	-1.122e-07	-9.592e-03	3.858e-02
12	-5.544e-08	-5.543e-08	1.877e-01	-6.532e-01
13	-2.490e-07	-2.843e-06	-1.126e-02	-8.674e-03
14	-7.634e-08	-7.634e-08	-1.039e-02	-9.719e-03
15	-1.884e-07	-5.871e-07	-2.825e-02	-3.754e-03
16	-1.007e-06	-1.007e-06	-5.842e-03	-1.452e-01

The smaller the absolute value in the table, the closer the corresponding cell probability is to the truth. Table 5 shows that compared with the simplest starting model $\mathscr{M}_{00}$, the model selected by the data-driven strategy has smaller absolute values of probability differences in almost all cells for the estimated U component, and in the majority of cells for the estimated M component, thus, leading to a better overall performance in this particular simulation case. This means that the selected model is closer to the true underlying model than the fitted $\mathscr{M}_{00}$. Similar patterns have been found in other simulation cases when the true model is $\mathscr{M}_{11}$. Therefore, the proposed strategy does have improvement in the model building compared

with the naive specification of field independence (i.e., starting model $\mathscr{M}_{00}$). Due to space limitation, we summarize only the cell probability differences between the selected and the true models, as well as those between the fitted $\mathscr{M}_{00}$ and the true model in Table 6 for U and M respectively, using the following measures for two general loglinear mixture models:

$$\begin{aligned} D(\mathscr{M}^{(0)}, \mathscr{M}^{(1)} \mid U) &= \sum_{i \in \text{all cells}} |p_i^{(0)}(U) - p_i^{(1)}(U)|, \\ D(\mathscr{M}^{(0)}, \mathscr{M}^{(1)} \mid M) &= \sum_{i \in \text{all cells}} |p_i^{(0)}(M) - p_i^{(1)}(M)|. \end{aligned}$$

Here $p_i^{(j)}$ is the probability of cell i in model $\mathscr{M}^{(j)}$. In addition, we also compare the estimated mixing proportion $\hat{\pi}(S)$ from the data-driven strategy and $\hat{\pi}(I)$ from the field independence specification with the true π: $\hat{\pi}(S) - \pi$ and $\hat{\pi}(I) - \pi$.

Table 6: Overall fitting comparisons between two fitted models and the true model.

Simulation Case	$D(S,T\|U)$	$D(I,T\|U)$	$D(S,T\|M)$	$D(I,T\|M)$	$\hat{\pi}(S)-\pi$	$\hat{\pi}(I)-\pi$
Setting 1, $\mathscr{M}_{00}$	1.668e-03	1.668e-03	1.967e-01	1.967e-01	-7.0e-06	-7.0e-06
Setting 2, $\mathscr{M}_{00}$	1.947e-03	1.947e-03	2.200e-01	2.200e-01	4.3e-05	4.3e-05
Setting 3, $\mathscr{M}_{00}$	2.059e-03	2.059e-03	3.287e-01	3.287e-01	-2.4e-05	-2.4e-05
Setting 4, $\mathscr{M}_{00}$	2.066e-03	1.487e-03	1.124e-01	1.009e-01	5.7e-06	-4.0e-06
Setting 5, $\mathscr{M}_{00}$	6.018e-04	6.018e-04	2.037e-01	2.037e-01	2.2e-04	2.2e-04
Setting 6, $\mathscr{M}_{00}$	5.984e-04	5.984e-04	1.293e-01	1.293e-01	1.2e-05	1.2e-05
Setting 1, $\mathscr{M}_{11}$	3.794e-04	2.400e-04	4.097e-01	1.482e-00	-1.5e-05	2.1e-04
Setting 2, $\mathscr{M}_{11}$	6.460e-04	7.581e-04	4.143e-01	5.985e-01	-3.0e-04	1.1e-04
Setting 3, $\mathscr{M}_{11}$	4.167e-04	7.049e-04	3.239e-01	1.103e-00	-4.5e-05	1.7e-04
Setting 4, $\mathscr{M}_{11}$	4.722e-04	4.678e-04	2.751e-01	3.372e-01	1.4e-04	4.6e-05
Setting 5, $\mathscr{M}_{11}$	7.136e-05	4.260e-04	3.755e-01	1.677e-00	-9.8e-07	2.6e-04
Setting 6, $\mathscr{M}_{11}$	2.453e-04	2.525e-04	4.145e-01	7.516e-01	-7.1e-05	1.1e-04

Table 6 shows that when the underlying true mixture model is $\mathscr{M}_{00}$, the data-driven strategy yields the same results as the Jaro's method if the selected model is consistent with $\mathscr{M}_{00}$. This means that the simplified method for the EM algorithm is equivalent to the Jaro's method when the field independence exists, and is thus an alternative approach. However, the advantage of this approach is that it can handle general contingency tables with factor levels more than two.

Regarding the performance of two fitted models in Table 6, when the field independence holds, the selected models by the data-driven strategy are almost the same as the fitted $\mathscr{M}_{00}$ except one overfitting case. However, they perform better than the fitted $\mathscr{M}_{00}$ when field independence is violated.

As to detecting interactions from the table data, our experience shows that it is not easy to capture the higher order interactions accurately. For this reason, we do not investigate further

the individual interactions in a given loglinear model in the proposed model building strategy, and it is also the partial reason why models are enlarged with a group of interactions in the building strategy in (11). Even more, for a data set simulated from a known loglinear model with three-factor interactions, the fitted model does not include these three-factor interactions although we increase the total count as many as possible in R. Therefore, in practice, it is likely to stop at a model with lower order interactions such as two-factor interactions, and this selected model may be a good approximation to the true model under the given sample size.

5. Application to a case study

We apply the proposed model building strategy to the record linkage case study at SSC 2006 (`www.ssc.ca/documents/case_studies/2006/record_index_e.html`).

The objective of the case study is to maintain a registry. Four completely synthetic data files named as "register", "births", "drivers" and "deaths" were constructed to simulate the registry for a sample of residents of the province of Prince Edward Island (PEI), Canada. The "register" file represents the population of PEI, containing a unique id, name, date of birth, and address information for residents. This is the file that we are working to maintain. The "births" file records new entrants into the population. They are persons who enter the population of interest, for example by moving into an area of interest, or attaining a certain age. The data file on births contains both present and previous address information as well as complete name and date of birth information. The "drivers" file provides moving information for registry. People living in Canada may take out their first license, get a new license (out of province) or update their present license (within a province). This file has information on name, address and date of birth. The "deaths" file contains name, address, date of birth and date of death. Note that data sources are updated independently. The main file is the "register" file which needs to be updated according to the other three files. In summary, the specific tasks include

(1) add new entrant records to "register" file from "births" file,

(2) update moving records to "register" file from "drivers" file,

(3) remove death records from "register" file according to "deaths" file.

To achieve these goals, we first need to identify common individuals in two compared files. Thus, record linkage is employed in these tasks.

Basically, the records in these files contain common personal information such as name (first, middle and last), date of birth (day, month and year), gender, address, and so on. We adopt the gender as the blocking factor to reduce both the number of record pairs and the number of factors in the fitted models. Table 7 summarizes the total pairs in each of three tasks after the gender blocking.

After a preliminary data analysis, we choose first name, middle name, last name and date of birth (day/month/year) as identifiers for the tasks of adding new entrant records and removing death records. For updating moving records, we only adopt first name, middle

Table 7: Numbers of record pairs between compared files after blocking using gender.

File 1	File 2	Gender	Pair number	Total Pair
register	births	Male	$2000 \times 102 = 204,000$	348,000
		Female	$2000 \times 72 = 144,000$	
register	drivers	Male	$2000 \times 470 = 940,000$	1968,000
		Female	$2000 \times 514 = 1,028,000$	
register	deaths	Male	$2000 \times 69 = 138,000$	330,000
		Female	$2000 \times 96 = 192,000$	

name, last name and postal code as identifiers, because the date of birth seems wrong by visual check. Furthermore, we exclude the records with missing values or error values on identifiers and gender in each file. These excluded records will be manually reviewed. The simple matching of "agreement" or "disagreement" is chosen as the comparison rule. Therefore, for the comparison vector $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3, \gamma_4)$, all four fields are binary, indicating match or non-match on identifiers. Each of the record pairs in Table 7 yields an observed comparison vector. Three contingency tables are then created for the comparison vectors, one for each task.

Next we apply the data-driven strategy to the three obtained contingency tables. The significance level is set as 5%. The selected models (in Poisson GLM formulation) are listed below.

(1) For adding new entrant records from the "births" file, γ_1, γ_2, γ_3 and γ_4 mean match or non-match on first name, middle name, last name and date of birth respectively. The fitted model is $\mathscr{M}_{10}$ with an estimated mixing proportion $\hat{\pi} = 4.885e-4$, where the U component is

$$\begin{aligned}&\alpha = 12.704,\ \alpha_1 = -6.586,\ \alpha_2 = -3.046,\ \alpha_3 = -5.998,\ \alpha_4 = -10.219,\\ &\alpha_{12} = 4.245,\ \alpha_{13} = 1.819,\ \alpha_{14} = -13.911,\ \alpha_{23} = 1.804,\ \alpha_{24} = 0.561,\\ &\alpha_{34} = -14.031,\end{aligned}$$

and the M component is

$$\alpha = -77.456,\ \alpha_1 = 27.503,\ \alpha_2 = 27.805,\ \alpha_3 = 0.332,\ \alpha_4 = 26.410.$$

(2) For updating moving records from the "drivers" file, γ_1, γ_2, γ_3 and γ_4 mean match or non-match on first name, middle name, last name and postal code respectively. The fitted model is $\mathscr{M}_{11}$ with an estimated mixing proportion $\hat{\pi} = 4.032e-4$, where the U component is

$$\begin{aligned}&\alpha = 14.439,\ \alpha_1 = -6.432,\ \alpha_2 = -3.128,\ \alpha_3 = -5.849,\ \alpha_4 = -8.882,\\ &\alpha_{12} = 4.205,\ \alpha_{13} = 1.427,\ \alpha_{14} = 1.569,\ \alpha_{23} = 1.879,\ \alpha_{24} = 0.427,\ \alpha_{34} = 1.807,\end{aligned}$$

and the M component is

$$\alpha = -21.502,\ \alpha_1 = -20.617,\ \alpha_2 = -5.895,\ \alpha_3 = 0.557,\ \alpha_4 = 0.249,$$

$$\alpha_{12} = 27.088,\ \alpha_{13} = 45.199,\ \alpha_{14} = -27.411,\ \alpha_{23} = -43.852,\ \alpha_{24} = 29.412,$$
$$\alpha_{34} = 23.367.$$

(3) For removing death records according to the "deaths" file, γ_1, γ_2, γ_3 and γ_4 mean match or non-match on first name, middle name, last name and date of birth respectively. The fitted model is $\mathscr{M}_{10}$ with an estimated mixing proportion $\hat{\pi} = 4.455e-4$, where the U component is

$$\alpha = 12.652,\ \alpha_1 = -6.725,\ \alpha_2 = -3.058,\ \alpha_3 = -5.531,\ \alpha_4 = -9.320,$$
$$\alpha_{12} = 3.794,\ \alpha_{13} = 2.159,\ \alpha_{14} = -20.402,\ \alpha_{23} = 1.391,\ \alpha_{24} = -0.302,$$
$$\alpha_{34} = -20.899,$$

and the M component is

$$\alpha = -28.828,\ \alpha_1 = 3.570,\ \alpha_2 = 3.871,\ \alpha_3 = 4.956,\ \alpha_4 = 21.366.$$

In these fitted models, we see interactions among names (first, middle and last), date of birth, and postal code. Note that these four synthetic files were actually constructed from other survey sources about PEI. So more or less they reflect the local information. The reason for the interactions among names may be traced back historically on the residents' family, origin and culture. In addition, PEI is a relatively closed island and the economy is not as active as most of other Canadian provinces. Families who are relatives might be more likely to live in a closed regions due to some reasons like traditional farming or other social activies, which could lead to the spatial patterns for surnames. Hence, these interactions are caused or confounded by the latent factors. Of course, these reasons could lead to higher order interactions, say three-factor interactions. However, with the current sample sizes, only two-factor interactions are significant enough to be included in the selected models.

We also build models for male and female block pairs separately in each task. Results are very similar and hence omitted.

It is straightforward to calculate matching weights and perform record pair partition for each task according to the corresponding selected model; this part is also omitted as it is not the focus of this paper.

6. Discussion

The simplified method for the EM algorithm is computationally effective. It also extends the Jaro's method for a general contingency table in fitting loglinear mixture models. This method can be easily implemented in a sophisticated statistical software using existing packages or functions for loglinear model fitting. We utilize R in this study. Compared with the Jaro's method, we extend not only the factor interactions but also the number of levels of each factor. The factor is not necessarily binary. It can be more than two levels for each factor.

This data-driven strategy of model building avoids subjective specification of the underlying loglinear mixture model, and thus minimizes the chance of underfitting or overfitting

caused by biased opinions. Often it yields a model consistent with the true one or close to the true one. Therefore, it is more likely to result in a better partition for record linkage. However, since the proposed strategy is not a global search, the selected model may not be the best.

Future investigations regarding other testing methods and building strategies, as well as error rate investigation are under way. We welcome any real case collaboration.

ACKNOWLEDGEMENTS This research has been supported by an NSERC Discovery grant. We thank Professor Román Viveros-Aguilera at McMaster University, Canada for his helpful discussion, and Miss Yunna Song for her participation as a team member in the record linkage case study at SSC 2006. In addition, we are grateful to a referee for comments leading to an improved presentation.

References

[1] E. D. Acheson. *Record Linkage in Medicine.* E. & S. Livingstone Ltd., Edinburgh and London, 1968.

[2] A. Agresti. *Categorical Data Analysis, Second Edition.* John Wiley & Sons, Inc., Hoboken, New Jersey, 2002.

[3] J. B. Copas and F. J. Hilton. Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society, A*, 153:287–320, 1990.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, B*, 39:1–38, 1977.

[5] M. E. Fair and P. Whitridge. Tutorial on record linkage. In W. Alvey and B. Jamerson, editors, *Record Linkage Techniques - 1997: Proceedings of an International Workshop and Exposition.*, pages 457–482, Arlington, VA, 1997. Federal Committee on Statistical Methodology, and Office of Management and Budget.

[6] I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64:1183–1210, 1969.

[7] M. A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa florida. *Journal of the American Statistical Association*, 84:414–420, 1989.

[8] M. A. Jaro. Probabilistic linkage of large public health datafiles. *Statistics in Medicine*, 14:491–498, 1995.

[9] M. D. Larsen. Modeling issues and the use of experience in record linkage. In W. Alvey and B. Jamerson, editors, *Record Linkage Techniques - 1997: Proceedings of an International Workshop and Exposition.*, pages 95–105, Arlington, VA, 1997. Federal Committee on Statistical Methodology, and Office of Management and Budget.

[10] M. D. Larsen and D. B. Rubin. Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96:32–41, 2001.

[11] G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley, New York, 1997.

[12] H. B. Newcombe. *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford University Press, Inc., New York, 1988.

[13] H. B. Newcombe, J. M. Kennedy, S. J. Axford, and A. P. James. Automatic linkage of vital records. *Science*, 130:954–959, 1959.

[14] Y. Thibaudeau. The discrimination power of dependency structures in record linkage. *Survey Methodology*, 19:31–38, 1993.

[15] W. E. Winkler. Using the em algorithm for weight computation in the fellegi-sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods.*, pages 667–671. American Statistical Association, 1988.

[16] W. E. Winkler. Method for adjusting for lack of independence in an application of the fellegi-sunter model of record linkage. *Survey Methodology*, 15:101–107, 1989.

[17] W. E. Winkler. Overview of record linkage and current research directions. Research Report Series Statistics #2006-2, U.S. Bureau of the Census, 2005.