Automatically Estimating Record Linkage False Match Rates

W. E. Winkler[1], william.e.winkler@census.gov
U.S. Bureau of the Census, Room 3000-4, Washington, DC 20233-9100

Abstract

This paper provides a mechanism for automatically estimating record linkage false match rates in situations where the subset of the true matches is reasonably well separated from other pairs and there is no training data. The method provides an alternative to the method of Belin and Rubin (*JASA* 1995) and is applicable in more situations. We provide examples demonstrating why the general problem of error rate estimation (both false match and false nonmatch rates) is likely impossible in situations without training data and exceptionally difficult even in the extremely rare situations when training data are available.

Keywords: EM algorithm, unsupervised and semi-supervised learning

1. Introduction

Record linkage is the science of finding matches or duplicates within or across files. Matches are typically delineated using name, address, and date-of-birth information. Other identifiers such as income, education, and credit information might be used. With a pair of records, identifiers might not correspond exactly. For instance, income in one record might be compared to mortgage payment size using a crude regression function.

In the model of record linkage due to Fellegi and Sunter (1969, hereafter FS), a product space $\mathbf{A} \times \mathbf{B}$ of records from two files A and B is partitioned into two sets *matches* M and *nonmatches* U. Pairs in M typically agree on characteristics (*quasi-identifiers*) such as first name, last name, components of date-of-birth, and address. Pairs in U often have isolated (random) agreements of the characteristics. We use $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n)$ to denote an arbitrary agreement pattern. For instance, γ might be agreement on first name, agreement on last name, and agreement on date-of-birth.

In the FS model, obtaining accurate estimates of the probabilities $P(\gamma \mid M)$ and $P(\gamma \mid U)$ are crucial to finding the best possible classification rules for separating matches M and nonmatches U. The conditional independence assumption **CI** is that $P(\gamma \mid C) = \prod_i P(\gamma_i \mid C)$ where the set C can be either M or U. Under **CI**, FS showed that it is possible to estimate $P(\gamma \mid M)$ and $P(\gamma \mid U)$ automatically without training data. For situations in which identifying information among matches is reasonably good, Winkler (1988) showed how to estimate $P(\gamma \mid M)$ and $P(\gamma \mid U)$ using the EM algorithm. The EM algorithm can provide good (sometimes optimal) separation between M and U because its parameters can correspond to the form needed for the classification rule. If assumption **CI** is not made, then a general EM (Winkler 1989, 1993, Larsen 1996) can provide parameters yielding better separation between M and U. The advantage of less general EM under assumption **CI** is that it yields computational speed-ups of orders between 100 and 1,000 in contrast to methods that use dependencies between variables. The disadvantage is that the **CI** EM yields probabilities that often do not correspond accurately to underlying truth at differing error-rate levels (Winkler 1993, Belin and Rubin 1995).

Winkler (2002, also Larsen and Rubin 2001) have demonstrated that, when small amounts of labeled data are combined with unlabeled data, parameter estimation and error-rate estimation can be improved. The relatively small samples are chosen among pairs where it is difficult to make a classification (i.e., pairs are chosen near decision rule boundaries).

The real world situation is that small amounts of labeled data that are a representative subset of the set of pairs being matched are almost impossible to determine. This is particularly true when matching parameters differ significantly across geographic regions, the minimal sample sizes of training data must be at least 500 pairs (less than 0.5% of pairs being matched), and the matching must be completed within a few days.

In this paper, we provide an unsupervised method for estimating false match rates. We do this for situations in which a subset of matches can be delineated with relatively high accuracy by a simplistic application of **CI** record linkage. We treat a subset of pairs above a certain score as 'pseudo-true' matches and a subset of pairs below another lower score as 'pseudo-true' nonmatches. With this artificial 'pseudo-truth' set, we apply the semi-supervised learning to all pairs under various modeling assumptions. In the best situation (i.e., most appropriate model), we obtain quite accurate estimates of the appropriate 30% tails of the curves of matches and nonmatches. These curves can, in turn, yield reasonably accurate estimates of false match rates in several situations and accurate estimates of false nonmatch rates.

[1] This report is released to inform interested parties of (ongoing) research and to encourage discussion (of work in progress).

The outline for this paper is as follows. In the second section, we cover background on the Fellegi-Sunter model, EM Algorithms and use of training data in semi-supervised learning situations. String comparators (Yancey 2005, Winkler 1990) break agreements of strings into five ranges corresponding to strong agreement, moderately strong agreement, weak agreement, missing (i.e., blank), and disagreement. The breakout of partial agreement into five subgroups contrasts to the agree/disagree or agree/disagree/blank groups of the earlier work. In the third section, we describe variants of the EM algorithm and the empirical data files on which we evaluate various models. The fourth section provides results. We give some discussion in the fifth section. The final section is concluding remarks.

2. Background

Fellegi and Sunter (1969) provided a formal mathematical model for ideas that had been introduced by Newcombe et al. (1959, 1962, see also 1988). They provided many ways of estimating key parameters. To begin, notation is needed. Two files **A** and **B** are matched. The idea is to classify pairs in a product space **A** × **B** from two files A and B into M, the set of true matches, and U, the set of true nonmatches. Fellegi and Sunter, making rigorous concepts introduced by Newcombe et al. (1959), considered ratios of probabilities of the form:

$$R = P(\gamma \in \Gamma \mid M) / P(\gamma \in \Gamma \mid U) \qquad (1)$$

where γ is an arbitrary agreement pattern in a comparison space Γ. For instance, Γ might consist of eight patterns representing simple agreement or not on the largest name component, street name, and street number. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific values of name components such as "Smith", "Zabrinsky", "AAA", and "Capitol" occur. The ratio R or any monotonely increasing function of it such as the natural log is referred to as a matching weight (or score).

The decision rule is given by:

If $R > T_\mu$, then designate pair as a match. (2a)

If $T_\lambda \le R \le T_\mu$, then designate pair as a possible match and hold for clerical review. (2b)

If $R < T_\lambda$, then designate pair as a nonmatch. (2c)

The cutoff thresholds T_μ and T_λ are determined by a priori error bounds on false matches and false nonmatches. Rule (2) agrees with intuition. If $\gamma \in \Gamma$ consists primarily of agreements, then it is intuitive that $\gamma \in \Gamma$ would be more likely to occur among matches than nonmatches and ratio (1) would be large. On the other hand, if $\gamma \in \Gamma$ consists primarily of disagreements, then ratio (1) would be small. Rule (2) partitions the set $\gamma \in \Gamma$ into three disjoint subregions. The region $T_\lambda \le R \le T_\mu$ is referred to as the no-decision region or *clerical review* region. In some situations, resources are available to review pairs clerically. Figure 1 provides an illustration of the curves of log frequency versus log weight for matches and nonmatches, respectively. The vertical lines in Figure 1 show hypothetical cutoff thresholds. The clerical review region (based on 1990 Decennial Census truth data) consists primarily of individuals in the same households that have missing values or severe errors in first name and age. Even small amounts of keypunch and transcription error can significantly affect typographical error and missing data.

Accurate estimation of error rates at different cutoff levels often depends on information that is not available. The most difficult part of the estimation is in the clerical review region that may have significantly varying characteristics in different geographic regions (such as between a suburban region and an adjacent urban region). These typographical error rates differ significantly depending on the type of region and how the computer files were pre-processed prior to computer matching. As an example, 0.01 percent of pairs in a suburban region with missing first name and age may be matches and 0.1 percent of pairs in an urban region with missing first name and age may be matches. Although these percentages of the clerical review region are quite small, the percentage of matches in these regions may be in the range 1-3% and substantially affect estimates of error rates.

A *false match* is a pair that is designated as a match and is truly a nonmatch. A *false nonmatch* is pair designated as a nonmatch and is a truly a match. If $\hat{M}$ are the pairs designated as matches by decision rule (2a), then the *false match rate* is given by $P(U \mid \hat{M})$. If $\hat{U}$ are the pairs designated as nonmatches by decision rule (2b), then $1 - P(M \mid \hat{U})$ is the *false nonmatch rate*.

3. Methods and Data

Our main theoretical method is to use the EM algorithm and maximum likelihood to obtain parameters and associated classifiers for separating A × B into matches M and nonmatches U. The data files are Decennial Census files for which the truth of

classification is known. The truth is obtained through several levels of clerical review, adjudication and field follow-up. The key difference with the earlier work (Winkler 2002) is that we 'artificially' designate a subset of pairs as 'pseudo-true' matches and 'pseudo-true' nonmatches. The earlier work needed to have small and moderate amounts of labeled data for which true matching status is known.

3.1 EM Methods

Our basic model is that of semi-supervised learning in which we combine a small proportion of labeled (true or pseudo-true matching status) pairs of records with a very large amount of unlabeled data. The conditional independence model corresponds to the naïve Bayesian network formulization of Nigam et al. (2000). The more general formulization of Winkler (2000, 2002) allows interactions between agreements (but is not used in this paper).

Our development is similar theoretically to that of Nigam et al. Our notation differs very slightly because it deals more with the representational framework of record linkage. Let γ_i be the agreement pattern associated with pair p_i. Classes C_j are an arbitrary partition of the set of pairs D in $\mathbf{A} \times \mathbf{B}$. Later, we will assume that some of the C_j will be subsets of M and the remaining C_j are subsets of U. Unlike general text classification in which every document may have a unique agreement pattern, in record linkage, some agreement patterns γ_i may have many pairs $p_{i(l)}$ associated with them. Here l will run through an appropriate index set. Specifically,

$$P(\gamma_i \mid \Theta) = \sum_i^{|C|} P(\gamma_i \mid C_j; \Theta)\, P(C_j ; \Theta) \qquad (3)$$

where γ_i is a specific pair, Cj is a specific class, and the sum is over the set of classes. Under the Naïve Bayes or conditional independence (**CI**), we have

$$P(\gamma_i \mid C_j ; \Theta) = \prod_k P(\gamma_{i,k} \mid C_j ; \Theta) \qquad (4)$$

where the product is over the k^{th} individual field agreement γ_{ik} in pair agreement pattern γ_i. In some situations, we use a Dirichlet prior

$$P(\Theta) = \Pi_j (\Theta_{Cj})^{\alpha-1} \quad \Pi_k (\Theta\, \gamma_{i,k \mid Cj})^{\alpha-1} \qquad (5)$$

where the first product is over the classes Cj and the second product is over the fields. We use Du to denote unlabeled pairs and Dl to denote labeled pairs. Given the set D of all labeled and unlabeled pairs, the log likelihood is given by

$$l_c(\Theta \mid D; z) = \log (P(\Theta)) +$$
$$(1-\lambda) \sum_{i \in Du} \sum_j z_{ij} \log (P(\gamma_i \mid C_j ; \Theta)\ P(C_j ; \Theta)) +$$
$$\lambda \sum_{i \in Dl} \sum_j z_{ij} \log (P(\gamma_i \mid C_j ; \Theta)\ P(C_j ; \Theta)). \qquad (6)$$

where $0 \le \lambda \le 1$. The first sum is over the unlabeled pairs and the second sum is over the labeled pairs. We observe that if λ is 1, then we only use training data and our methods correspond to naïve Bayes methods in which training data are available. If λ is 0, then we are in the unsupervised learning situations of Winkler (1993) and Larsen (1996). Winkler (2002, 2000) provides more details of the computational algorithms.

3.2 Data Files

Three pairs of files were used in the analyses. The files are from 1990 Decennial Census matching data in which the entire set of 1-2% of the matching status codes that were believed to have been in error for these analyses have been corrected. The corrections reflect clerical review and field follow-up that were not incorporated in computer files available to us.

A summary of the overall characteristics of the empirical data is in Table 2. We only consider pairs that agree on census block id (small geographic area representing approximately 70 households) and on the first character of surname. Less than 1-2% of the matches are missed using this set of blocking criteria. They are not considered in the analysis of this paper.

Table 2. Summary of Three Pairs of Files

	Files $A1_1$ $A2_2$	Files $B1_1$ $B2_2$	Files C1 $C2_2$
Size	15048 12072	4539 4851	5022 5212
# pairs	116305	38795	37327
# matches	10096	3490	3623

The matching fields that are:

Person Characteristics: First Name, Age, Marital Status, Sex
Household Characteristics: Last Name, House Number, Street Name, Phone

Typically, everyone in a household will agree on the household characteristics. Person characteristics such as first name and age help distinguish individuals within household. Some pairs (including true matches) have both missing first name and age.

We also consider partial levels of agreement in which the string comparator values are broken out as [0, 0.66], (0.66, 0.88], (0.88, 0.94], and (0.94, 1.0]. The first interval is what we refer to as disagreement. We combine the disagreement with the three partial agreements and blank to get five value states (base 5).

The large base analyses consider five states for all characteristics except sex and marital status for which we consider three (agree/blank/disagree). The total number of agreement patterns is 140,625. In the earlier work (Winkler 2002), the five levels of agreement worked consistently better than two levels (agree/disagree) or three levels (agree/blank/disagree).

The pairs naturally divide into three classes: C_1-match within household, C_2 - nonmatch within household, C_3 – nonmatch outside household. In the earlier work (Winkler 2002), we considered two dependency models in addition to the conditional independence model. In that work in which small amounts of labeled training data were combined with unlabeled data, the conditional independence model worked well and the dependency models worked slightly better.

Metaparameters (Table 3) are generally decided prior to fitting various types of models. Generally the iterations of the fitting are less than 100. The delta value allows probability mass in cells where the observed population of pairs may have a (sampling) zero.

Table 3. Metaparameters of the Modeling

1. Model – **CI** – independent
2. lambda – how much to emphasize training data
3. delta – 0.000001 to 0.001 – smooth out peaks
4. how many iterations
5. number of degrees of partial agreement
   very close agree, moderately close agree, somewhat agree, blank, disagree [large base =5]

For comparison with and understanding of the previous work (Winkler 2002), we need training samples of labeled data for which true matching status is known. The samples are concentrated in the clerical review regions which are between the vertical bars of Figure 1. In clerical review regions, it is difficult to distinguish between matches and nonmatches (see section 2). We draw relatively small and relatively large samples of training data. The sample sizes are summarized in Table 4.

Table 4. Training Data Counts with Proportions of Matches for Earlier Work (Winkler 2002)

Sample	A	B	C
Large	7612 (0.26)	3031 (0.29)	3287 (0.27)
Small	588 (0.33)	516 (0.26)	540 (0.24)

Under each of the scenarios, we do semi-supervised learning (λ = 0.9 or 0.99). In the semi-supervised learning situation, we use both large and small samples that are concentrated in regions where it is difficult to make a decision between match and nonmatch.

We create 'pseudo-truth' data sets in which matches are those unlabeled pairs above a certain high cutoff and nonmatches are those unlabeled pairs below a certain low cutoff. Figure 1 illustrates the situation using actual 1990 Decennial Census data in which we plot log of the probability ratio (1) against the log of frequency. With the datasets of this paper, we choose high and low cutoffs in a similar manner so that we do not included in-between pairs in our designated 'pseudo-truth' data sets. We use these 'designated' pseudo-truth data sets in a semi-supervised learning procedure that is nearly identical to the semi-supervised procedure where we have actual truth data. A key difference from the corresponding procedure with actual truth data is that the sample of labeled pairs is concentrated in the difficult-to-classify in-between region where, in the 'pseudo-truth' situation, we have no way to designate comparable labeled pairs. The sizes of the 'pseudo-truth' data is given in Table 5. The errors associated with the artificial 'pseudo-truth' are given in parentheses following the counts. The *Other* class gives counts of the pairs and proportions of true matches that are not included in the 'pseudo-truth' set of pairs. In the *Other* class, the proportions of matches vary somewhat and would be difficult to determine without training data.

Table 5. 'Pseudo-Truth' Data with Actual Error Rates

	Matches	Nonmatches	*Other*
A pairs	8817 (.008)	98257 (.001)	9231 (.136)
B pairs	2674 (.010)	27744 (.0004)	8377 (.138)
C pairs	2492 (.010)	31266 (.002)	3569 (.369)

Our empirical results are only for the non-1-1 matching situation. We determine how accurately we can estimate the lower cumulative distributions of matches and the upper cumulative distribution of nonmatches. This corresponds to the overlap region of the curves of matches and nonmatches. If we can accurately estimate these two tails of distributions, then we can accurately estimate error rates at differing levels. Our comparisons consist of a set of figures in which we compare a plot of the cumulative distribution of estimates of matches versus the true cumulative distribution with the truth represented by the 45 degree line. We also do this for nonmatches. As the plots get closer to the 45 degree lines, the estimates get closer to the truth.

In the earlier work (Winkler 2002) in which small amounts of training data were needed and combined with large amounts of unlabeled data, we showed that base 5 results were uniformly better than base 2 results, that error-rate estimates were quite accurate under

conditional independence, and slightly more accurate under the dependency models. Because training data are typically not available and optimal parameters vary significantly across regions (Winkler 1989), we investigate the methods of this paper to determine whether it is possible to obtain reasonably accurate estimates of error rates without training data.

4. Results

Our primary results are from using the conditional independence model and 'semi-supervised' methods of this paper with the conditional independence model and actual semi-supervised methods of Winkler (2002). With our 'pseudo-truth' data, we obtain the best sets of estimates of the bottom 30% tails of the curve of matches and the top 5% tails of nonmatches with conditional independence and λ=0.2. Figure 2a-f illustrates the set of curves that provide quite accurate fits. The 45 degree line represents the truth whereas the curve represents the cumulative estimates of matches and nonmatches for the left and right tails, respectively. Although we looked at results for λ=0.1, 0.5, and 0.8 and various interactions models, the results under conditional independence (**CI**) were the best with λ=0.2. We also looked at several different ways of constructing the 'pseudo-truth' data.

In Figure 3a-f, we provide estimates of the appropriate cumulative tails of the false match rates and false nonmatch rates, respectively. We observe that, with the exception of the false match curve for file-set C, estimates are reasonably accurate. We have no explanation for why the false match curve for file-set C shows substantial inaccuracy. Figures 4a-f provide comparison with previous work (Winkler 2002) for the small sample situation with λ=0.99. The small samples situation provides slightly more accurate estimates of error rates than the estimation procedure of this paper in which no labeled training data is available. This is particularly true in the extreme tails of the distributions where the labeled data from the in-between region provides much better information than the situation of this paper where we have no comparable information to use during the estimation procedure.

5. Discussion

In this section, we describe the reasons for the general inability to estimate false nonmatch rates accurately with most pairs of lists. We could estimate false non-match rates for the pairs of files of this paper because of the relatively higher quality of name and other information in comparison with general files and because field follow-up, adjudication, and additional review was used to located virtually all matches.

There are many pairs of files for which the quasi-identifying information (names, address, dates-of-birth, etc.) are very different or completely different for the true matches that we wish to match. In these situations, we cannot effectively bring the pairs together and it is (nearly) impossible to determine how many matches we have missed.

The most straightforward situation is illustrated in Table 6. The name information in the first file is out-of-date because Susan K. Jones has changed her legal name to Susan K. Smith and she usually uses her middle name Karen (File B). The date-of-birth in the first file is correct whereas the date-of-birth in the second file is completely wrong.

Table 6. Two Name and Date-of-Births associated with the Same Individual

	Name	Date-of-birth
File A	Susan K. Jones	January 16, 1964
File B	Karen Smith	November 10, 1975

Some person lists may have substantially more of the difficulties of the type illustrated in Table 6 than other person lists. As a particular instance, one keypuncher who learns how to override date-of-birth keypunch controls can assure that most (or all) of dates-of-birth are in error in the records that he/she has keyed.

With many business lists, the representations of the names (John K Smith and Company versus JKS Inc) and the mailing addresses (street address of physical location versus PO Box) may be completely different. Business lists always have substantially higher proportions of name and address representation difficulties than person lists. In each type of situations, it is impossible to determine whether a small, moderate, or proportion of the pairs should have been matches and were not detected without auxiliary information.

We can also have difficulties locating matches automatically in situations where a modest proportion of records in a file fail name standardization, fail address standardization, or have dates-of-birth that cannot be effectively put in the form of dates-of-birth in a file to which the first file is being compared.

6. Concluding Remarks

In earlier work (Winkler 2002), we demonstrated, if small amounts of properly chosen training data are combined with large amounts of unlabeled data (i.e., semi-supervised learning), then we could obtain accurate estimates of error rates in a variety of situations. Because labeled training data are often unavailable, we provide an unsupervised learning method (no training data) that works almost as well as the previous method.

References

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

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

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

Larsen, M. D., and Rubin, D. B. (2001), "Iterative Automated Record Linkage Using Mixture Models," *Journal of the American Statistical Association*, 79, 32-41.

Newcombe, H. B. (1988), *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*, Oxford: Oxford University Press.

Newcombe, H. B., Kennedy, J. M. Axford, S. J., and James, A. P. (1959), "Automatic Linkage of Vital Records," *Science*, 130, 954-959.

Newcombe, H.B., and Kennedy, J. M. (1962) "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information" *Communications of the Association for Computing Machinery*, .5, 563-567.

Nigam, K., McCallum, A. K., Thrun, S. and Mitchell, T. (2000), "Text Classification from Labeled and Unlabelled Documents using EM, *Machine Learning*, 39, 103-134.

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

Winkler, W. E. (1989), "Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Fifth Census Bureau Annual Research Conference*, 145-155.

Winkler, W. E. (1990), "On Dykstra's Iterative Fitting Procedure," *Annals of Probability*, 18, 1410-1415.

Winkler, W. E. (1990), "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 354-359.

Winkler, W. E. (1993), "Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 274-279.

Winkler, W. E. (1994), "Advanced Methods for Record Linkage," *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 467-472 (longer version report 94/05 available at http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. (2000), "Machine Learning, Information Retrieval, and Record Linkage," *American Statistical Association*, *Proceedings of the Section on Survey Research Methods*, 20-29. (also available at http://www.niss.org/affiliates/dqworkshop/papers/winkler.pdf).

Winkler, W. E. (2002), "Record Linkage and Bayesian Networks," *American Statistical Association*, *Proceedings of the Section on Survey Research Methods*, CD-ROM (also report RRS2002/05 at http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. (2004), "Approximate String Comparator Search Strategies for Very Large Administrative Lists," *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, CD-ROM (also report 2005/06 at http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. (2006), "Overview of Record Linkage and Current Research Directions," Statistical Research Division Research Report, http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf.

Yancey, W. E. (2005), "Evaluating String Comparator Performance for Record Linkage," Statistical Research Division Research Report, http://www.census.gov/srd/papers/pdf/rrs2005-05.pdf.

Figure 1. Log Frequency vs Weight
Matches and Nonmatches Combined

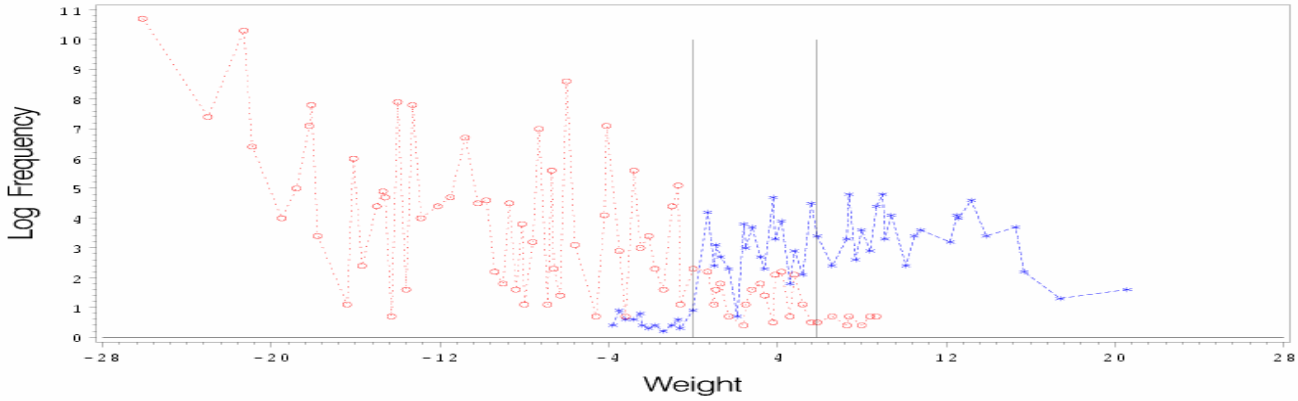

Figure 2a. Estimates vs Truth, File A
Cumulative Matches, Tail of Distribution
Independent EM, Lambda = 0.2

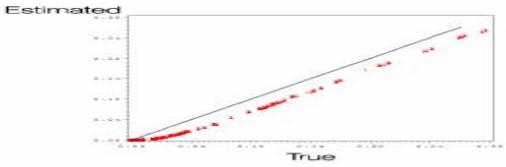

Figure 2b. Estimates vs Truth, File A
Cumulative Nonmatches, Tail of Distribution
Independent EM, Lambda = 0.2

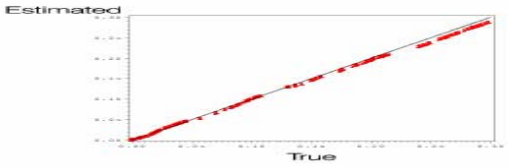

Figure 2c. Estimates vs Truth, File B
Cumulative Matches, Tail of Distribution
Independent EM, Lambda = 0.2

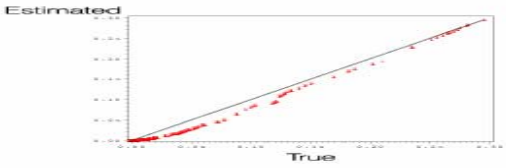

Figure 2d. Estimates vs Truth, File B
Cumulative Nonmatches, Tail of Distribution
Independent EM, Lambda = 0.2

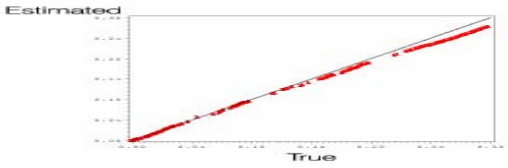

Figure 2e. Estimates vs Truth, File C
Cumulative Matches, Tail of Distribution
Independent EM, Lambda = 0.2

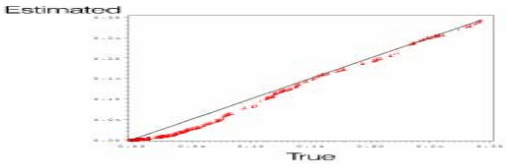

Figure 2f. Estimates vs Truth, File C
Cumulative Nonmatches, Tail of Distribution
Independent EM, Lambda = 0.2

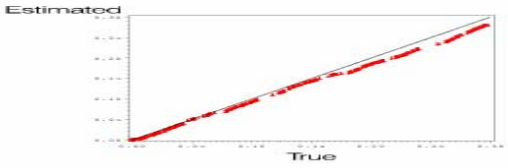

Figure 3a. Estimates vs Truth, File A
Cumulative False Match Rates by Weight
Independent EM, Lambda = 0.2

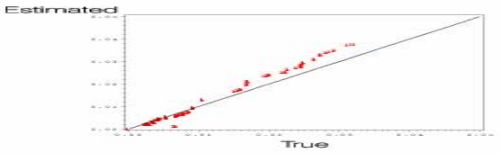

Figure 3b. Estimates vs Truth, File A
Cumulative False Nonmatches by Weight
Independent EM, Lambda = 0.2

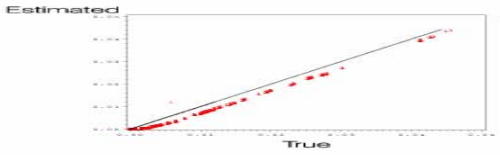

Figure 3c. Estimates vs Truth, File B
Cumulative False Match Rates by Weight
Independent EM, Lambda = 0.2

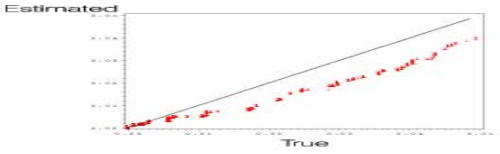

Figure 3d. Estimates vs Truth, File B
Cumulative False Nonmatches by Weight
Independent EM, Lambda = 0.2

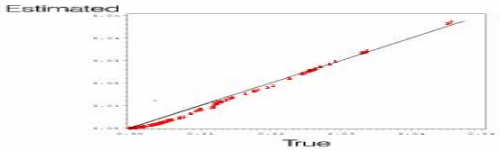

Figure 3e. Estimates vs Truth, File C
Cumulative False Match Rates by Weight
Independent EM, Lambda = 0.2

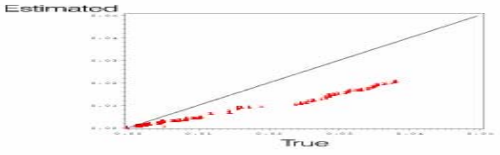

Figure 3f. Estimates vs Truth, File C
Cumulative False Nonmatches by Weight
Independent EM, Lambda = 0.2

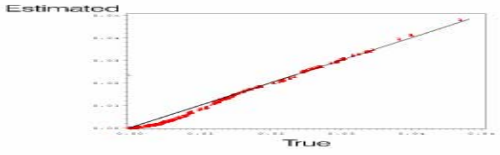

Figure 4a. Estimates vs Truth, File A
Cumulative Matches, Lambda = 0.99
Small Sample, Independent EM, non-1-1

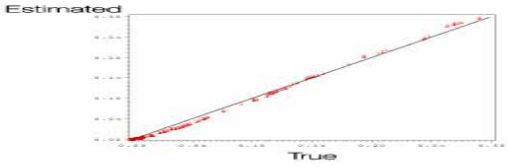

Figure 4b. Estimates vs Truth, File A
Cumulative Nonmatches, Lambda = 0.99
Small Sample, Independent EM, non-1-1

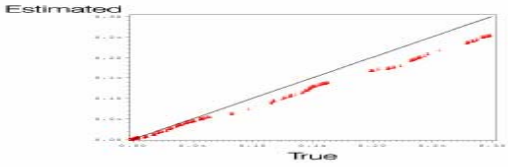

Figure 4c. Estimates vs Truth, File B
Cumulative Matches, Lambda = 0.99
Small Sample, Independent EM, non-1-1

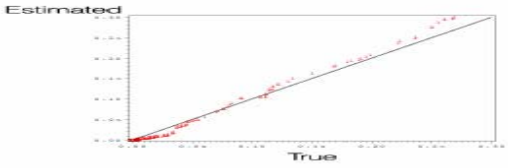

Figure 4d. Estimates vs Truth, File B
Cumulative Nonmatches, Lambda = 0.99
Small Sample, Independent EM, non-1-1

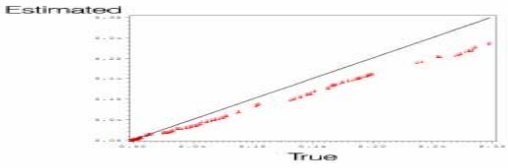

Figure 4e. Estimates vs Truth, File C
Cumulative Distribution of Matches, Lambda = 0.99
Small Sample, Independent EM, non-1-1

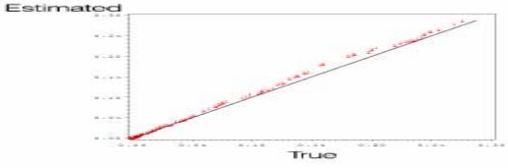

Figure 4f. Estimates vs Truth, File C
Cumulative Nonmatches, Lambda = 0.99
Small Sample, Independent EM, non-1-1