Methods for Record Linkage and Bayesian Networks

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

ABSTRACT

Although terminology differs, there is considerable overlap between record linkage methods based on the Fellegi-Sunter model (JASA 1969) and Bayesian networks used in machine learning (Mitchell 1997). Both are based on formal probabilistic models that can be shown to be equivalent in many situations (Winkler 2000). When no missing data are present in identifying fields and training data are available, then both can efficiently estimate parameters of interest. EM and MCMC methods can be used for automatically estimating parameters and error rates in some of the record linkage situations (Belin and Rubin 1995, Larsen and Rubin 2001).

Keywords: likelihood ratio, Bayesian Nets, EM Algorithm

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 computer science literature, *data-cleaning* often refers to methods of finding duplicates.

In the model of record linkage due to Fellegi and Sunter (1969, hereafter FS), a product space **A** × **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 such as first name, last name, components of date-of-birth, and address. Pairs in U typically 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 decision 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 is useful because it provides a means of optimally separating M and U. Better separation between M and U is possible with a general EM (Winkler 1989, 1993, Larsen 1996) that does not use assumption CI.

Bayesian Networks are graphical networks that are often used in the machine learning literature. A Naïve Bayes Network is a Bayesian Network under assumption CI. Naïve Bayes networks are typically applied in situations in which representative training data are available. Naïve Bayes methods have been extended to situations in which a mixture of labeled training data and unlabeled data are used for text classification (Nigam et al. 2000). Parameter estimation was done using a version of the EM algorithm that is effectively identical to that used by Winkler (2000) and Larsen and Rubin (2001) when training data are not available. In the latter situations, assumption CI was not needed.

Nigam et al. (2000) demonstrated if a small amount of labeled training data is combined with a moderate or relatively large amount of unlabelled data, then classification rules are improved. The improvement is in contrast to those methods in which only labeled data are used in training. In most situations for machine learning, training data provides necessary structure so that parameter estimation can provide classification rules that perform relatively well. In contrast to unsupervised learning methods for which no training data are available, the training data drastically reduces the number of computational paths that are considered by the parameter-estimation algorithms.

The unsupervised learning methods of record linkage (Winkler 1988, 1993) performed relatively well because they were applied in a few situations that were extremely favorable. Five conditions are favorable application of the unsupervised EM methods. The first is that the EM must be applied to sets of pairs in which the proportion of matches M is greater than 0.05 (see Yancey 2002 for related work). The second is that one class (matches) must be relatively well-separated from the other classes. The third is that typographical error must be relatively low. For instance, if twenty percent of matches have first name pairs that are of the form (Robert, brother), (Bob, blank), or (Robert, James) then it may be difficult to separate matches from nonmatches. The fourth is that there must be redundant identifiers that overcome errors in other identifiers. The fifth is that parameters obtained under assumption CI yield good classification rules. Under the five favorable conditions, the number of computational paths considered by the EM algorithm is greatly reduced from the number of computational paths under general EM when the five assumptions are known not to hold.

This paper will be concerned with discovering situations when some of the five assumptions can be weakened. We investigate more complicated models than those with condition CI, simple application of string comparators, and use of training data. Within this framework, it is possible to get better parameter estimates and better record linkage classification rules. The main intent is to focus on relatively parsimonious computational extensions of the narrowest EM methods. One goal is a method for improving matching efficacy in very large administration list situations when each list may contain between 100 million and 1 billion records. The improvements reduce clerical review regions and allow estimation of error rates.

The outline for this paper is as follows. In the second section, we cover background on the Fellegi-Sunter model, Bayes Networks, EM Algorithms, use of training data, and effects of typographical error in identifiers. In the third section, we describe variants of the EM algorithm and the empirical data files. The fourth section provides results. We give some discussion in the fifth section. The final section is concluding remarks.

2. BACKGROUND

This section gives basic background on record linkage and specific issues that relate to it. In the first subsection, we formally describe the main theory of Fellegi and Sunter. The second subsection covers Bayesian Networks and their relationship to Fellegi-Sunter theory. The third subsection specifically gives insights into the strengths and limitations of training data in the record linkage setting. The fourth subsection give reasons why typographical error and related representational differences affect and limit efficacy of EM methods. The fifth subsection describes how different sets of blocking criteria and specific ways of applying weak classifiers create suitable sets of pairs for later applying stronger classifiers for separating matches from nonmatches. In the sixth subsection, we cover extensions for approximate string comparison.

2.1. Fellegi-Sunter Model of Record Linkage

Fellegi and Sunter (1969) provided a formal mathematical model for record linkage. 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 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.

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

If $R < T_\lambda$, then designate pair as a nonmatch.

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.

Pairs with weights above the upper cut-off are referred to as *designated matches* (or links). Pairs below the lower cut-off are referred to as *designated nonmatches* (or nonlinks). The remaining pairs are referred to as *designated potential matches* (or potential links). If $T_\mu = T_\lambda$, then decision rule (1) can be used for separating records (correspondingly pairs) into those that are in one class from those that are not. The probabilities P(agree first | M), P(agree last | M), P(agree age | M), P(agree first | U), P(agree last | U), and P(agree age | U) are called *marginal probabilities*. P(| M) & P(| U) are called the m- and u-probabilities, respectively. The natural logarithm of the ratio R of the probabilities is called the *matching weight or total agreement weight*. The logarithms of the ratios of probabilities associated with individual fields (marginal probabilities) are called the *individual agreement weights*. The m- and u-probabilities are also referred to as *matching parameters*. 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.

2.2. Bayesian Networks

Nigam et al. (2000) observed two strengths of Bayesian networks. The first is that the method is based on a formal probabilistic model that lends itself to statistical interpretation. The second is that it provides a straightforward way of combining labeled and unlabelled data during training. In most machine learning applications, only labeled training data are used. Because training data are very expensive and unlabelled data are easy to collect, Nigam et al. (2000) showed how to combine moderate amounts of labeled data with varying amounts of unlabelled data to produce classification decision rules that improved on classification rules that were based on the moderate amounts of labeled data alone.

Nigam et al. (2000) have shown that classification decision rules that are based on naïve Bayesian networks (i.e., conditional independence assumption) work well in practice. The conditional independence is useful because it makes computation much more tractable (Nigam et al. 2000, Winkler 1988). Varying authors have observed that the fields in pairs are quite dependent and that the computed probabilities for pairs do not even remotely correspond to the true underlying probabilities. Winkler (1989, 1993) observed that, if dependencies are dealt with, computed probabilities can somewhat correspond to the true probabilities in a few situations. Dependencies can be computed with conventional hierarchical latent class methods as in Winkler (1989, 1993) when the number of fields is moderate (say, 20 or less) or in much larger problems (Winkler 2000).

2.3. Use of Training Data

For general machine learning, it is well known that a suitable amount of representative data can yield good classifiers. A suitable amount of training data is generally O(m) where m are the number of parameters in the machine learning. When m is moderately large, then the amount of training data O(m) can be quite large and expensive to obtain. The intuition of using a more moderate amount of training data with a relatively large amount of (inexpensive) unlabelled data is two-fold. First, the training data provides more structure for the computational algorithms in the sense that they can severely limit the number of computational paths. Second, when the unlabelled data is combined in a suitable manner with labeled training data, effective size of training data is increased. In some situations, Larsen and Rubin (2001) showed that training with unlabeled data and increasing amounts of labeled training data can significantly improve matching efficacy.

2.4. Errors in Identifiers

Record linkage and general text classification have a major difference because, in record linkage, we need nearly automatic methods of dealing with differing amounts of typographical error in pairs of files. For instance, P(agree characteristic | M) and P(agree characteristic | U) can very dramatically from one pair of files to another. More surprising is that these probabilities can vary substantially from an urban region to an adjacent suburban region even when identical sets of fields are used in the matching. Fellegi and Sunter (1969) and Winkler (1995) indicated that part (or most) of the difference is due to differing amounts of typographical error.

2.5. Identifying Suitable Sets of Pairs

It is not possible to consider all pairs from two files A and B. In record linkage, we consider only those pairs agreeing on blocking criteria. One blocking pass might consider only those pairs agreeing on a geographic identifier and the first character of surname. The idea of blocking is to find a set of pairs in which matches are concentrated. Multiple blocking passes are needed to find duplicates in a subsequent blocking pass that are not found on a prior pass.

Unlike general text classification, in record linkage it is quite feasible to use an initial guess of parameters associated with agreements to get an enriched set of pairs within a blocking criteria. Virtually all matches will be concentrated in those pairs having matching weight (1) above a certain value. Yancey (2002) shows how to improve matching parameters within such classes of pairs via an EM algorithm.

2.6. Approximate String Comparison

Many matches have typographical error in key identifying fields. For instance, in 1988 Dress Rehearsal Census data among pairs that are true matches, 20% of first names and 15% of last names did not agree on an exact character-by-character basis. Ages were missing differed by more than 1 year with at least 15%. To alleviate some of the effect of typographical error, we use string comparators that return values between 1 for exact agreement and 0 for total disagreement (e.g.. Winkler 1995).

3. METHODS AND DATA

Our main theoretical method is to use the EM algorithm 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 via clerical review, field followup and adjudication.

3.1. EM Methods

In this section, we very briefly summarize the EM parameter-estimation method. A more complete development is in Nigam et al. (2000) and Winkler (2000, 1993). Our development is identical theoretically to that of Nigam et al. 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 **A** × **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. 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, C_j is a specific class, and the sum is over the set of classes. In some situations, we use a Dirichlet prior

$$P(\Theta) = \prod_j (\Theta_{Cj})^{\alpha-1} \quad \prod_k (\Theta \gamma_{i,k \mid Cj})^{\alpha-1} \qquad (4)$$

where the first product is over the classes C_j and the second product is over the fields. The prior (4) helps keep most of the estimated probabilities away from zero. We use D^u to denote unlabeled pairs and D^l to denote labeled pairs. Given the set D, if we have labeled and unlabelled pairs are mixed in proportions λ and $1-\lambda$, $0< \lambda < 1$, and let z_{ij} be a missing data indicator that pair i in class j is observed, then we have the complete data equation (CDE) log likelihood is given by

$$\begin{aligned} l_c(\Theta \mid D; z) = &\log (P(\Theta)) + \\ &(1-\lambda) \sum_{i\in Du} \sum_j z_{ij} \log (P(\gamma_i \mid C_j ; \Theta) \; P(C_j ; \Theta)) + \\ &\lambda \sum_{i\in Dl} \sum_j z_{ij} \log (P(\gamma_i \mid C_j ; \Theta) \; P(C_j ; \Theta)). \qquad (5) \end{aligned}$$

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 followup that were not incorporated in computer files available to us.

A summary of the overall characteristics of the empirical data is in Table 1. We only consider pairs that agree on census block id and on the first character of surname. At most 1-2% of the matches are missed by using this blocking criteria. They are not considered in the analysis of this paper.

Table 1. Summary of Three Pairs of Files

	Files		Files		Files	
	A_1	A_2	B_1	B_2	C_1	C_2
Size	15048	12072	5022	5212	4539	4851
# pairs	116305		37327		38795	
# matches	10096		3623		3490	

The matching fields that are:

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

Typically, everyone in a household will agree on the household characteristics. Person characteristics help distinguish individuals within household. Some pairs have both missing first name and age. In the initial results, all comparisons are considered agree/disagree (base 2). This basic situation corresponds to matching comparisons that were used in matching systems in 1990 and 2000. The eight data fields yield 256 data patterns for which frequencies are calculated. If one or both identifiers of a pair are blank, then the comparison (blank) is considered a disagreement. This only substantially affects age (5-10% blank) and phone (30% blank). Name and address data are almost never missing.

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]. 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 3 (agree/blank/disagree). The total number of agreement patterns is 140,625.

The pairs naturally divide into three classes: C_1- match within household, C_2 - nonmatch within household, C_3 – nonmatch outside household. Although we considered additional dependency models, we present results for only two models. The first is of Larsen and Rubin (2001), called g_1: I,HP,HP, in which we fit a conditional independence model in class C_1 and 4-way interaction models in classes C_2 and C_3. The second is similar to ones considered by Winkler (1993). It is called g_3: HP+,HP+,HP+ in which we fit slightly more interactions than in g_1 in all three classes. The analysis framework is quite flexible. It is summarized in Table 2. We refer to the different five components in Table 2 as the metaparameters of the modeling framework. This is reasonably consistent with Hastie, Thibshirani, and Friedman (2001).

Table 2. Metaparameters of the Modeling

1. Models – CI – independent – *i1(i0 – 1990 version) I,I,I*
   Larsen-Rubin CI in class 1, 4-way person,
   4-way household in classes 2 and 3, *g1 I,HP,HP*
   Winkler 4+ way interactions in all classes,
   g3(g0 1990 version)
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
   a. agree, disagree (and/or blank) [small base = 2]
   b. very close agree, moderately close agree, somewhat agree, blank, disagree [large base = 5]

We draw small and relatively large samples of training data. The sample sizes are summarized in Table 3.

Table 3. Training Data Counts with Proportions of Matches

	A	B	C
Large Sample	7612 (.26)	3031 (.29)	3287 (.27)
Small Sample	588 (.33)	516 (.26)	540 (.24)

The overall comparisons are summarized in Table 4. Under each of the scenarios, we do unsupervised learning (λ <= 0.001) and supervised learning (λ = 0.9, 0.99 or 0.999). In the supervised learning situation, we use both large and small samples.

Table 4. Summary of Comparison Scenarios

1990	2002
yes/no	3-level yes, blank, no
CI (i0), interact (g0)	CI (i1), interact (g1, g3)
I,I,I ; HP+,HP+,HP+	I,I,I ; I,HP,HP; HP+,HP+,HP+
1-1, non-1-1	1-1, non-1-1
no delta	delta smoothing

We have two overall measures of success. The first is applied only when we use 1-1 matching. At a set of fixed error levels (0.002, 0.005, 0.01, and 0.02), we consider the number of pairs brought together and the proportion of matches that are obtained. This corresponds to production matching systems used in 1990 and 2000. The second is applied only when we use non-1-1 matching. We determine how accurately we can estimate the lower cumulative distributions 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. This is known to be an exceptionally difficult problem (e.g., Hastie, Thibshirani, and Friedman 2001). 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. The closer the plots are to the 45 degree lines the closer the estimates are to the truth.

4. RESULTS

The results are divided into two subsections. In the first, we consider results from 1-1 matching. In the second, much harder situation, we consider the estimation of the tails of distributions.

4.1. Results under 1-1 Matching

Table 5 gives results from 1-1 matching. At differing error rate levels and in the differing files, the 1990 matching procedures that were also used in 2000 were nearly as effective as the newer procedures. Use of the larger base sometimes improves results by 0.005. Use of interaction models sometimes improves results by 0.005.

4.2. Results under non-1-1 Matching

Although we looked at a very large number of scenarios using different mixing proportions, interaction patterns and bases, our results are summarized in Figures 1 and 2. Figure 1 shows that the best method of unsupervised learning using the conditional independence assumption, large-base and 1-1 matching. It improves slightly on the unsupervised learning methods of Winkler (1988). It does not provide suitable error-rate estimates in the crucial 0.0-0.1 range. Results (not shown) with the interaction models and 1-1 matching were worse. Figure 2 shows results under the conditional independence assumption and with a small training sample (approximately 0.0001 of the pairs). It provides suitable error-rate estimates in the 0.0-0.1 range. Results (not shown) for small samples and interactions models were worse. The small sample size data did not contain sufficient information. Results (not shown) with large samples were uniformly better with interaction models performing best.

Table 5. Matching efficacy for 1-1 matching
Number of pairs with proportions of true matches

Error level	File A	File B	File C
0.002			
g3	9780 (0.967)	3428 (0.944)	3225 (0.922)
g1	9741 (0.965)	3448 (0.950)	3261 (0.932)
i1	9640 (0.956)	3277 (0.903)	3042 (0.867)
i0	9701 (0.959)	3489 (0.961)	3306 (0.945)
g0	9649 (0.954)	3422 (0.943)	3273 (0.936)
0.005			
g3	9882 (0.974)	3547 (0.974)	3409 (0.972)
g1	9868 (0.973)	3523 (0.967)	3386 (0.965)
i1	9855 (0.971)	3513 (0.965)	3314 (0.945)
i0	9857 (0.971)	3540 (0.972)	3379 (0.963)
g0	9810 (0.967)	3511 (0.964)	3329 (0.949)

5. DISCUSSION

See longer research report.

6. CONCLUDING REMARKS

This paper examines methods for weakening some of the stringent conditions that were in earlier applications of EM parameter estimation to record linkage. The EM-based estimation methods yield better parameters for separating matches from nonmatches. The estimates improve over iterative refinement methods that proceed through a series of clerical reviews and expert guessing such as are available in certain commercial record linkage systems. They are also far faster and better use resources than iterative refinement methods.

1/ This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion. A longer research report is available at http://www.census.gov/srd/www/byyear.html.

REFERENCES

See longer research report.

Figure 1a. Estimates vs Truth, File A
Cumulative Distribution of Matches
Large base, Independent EM, non−1−1

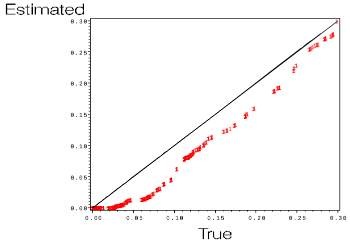


Figure 1c. Estimates vs Truth, File B
Cumulative Distribution of Matches
Large base, Independent EM, non−1−1

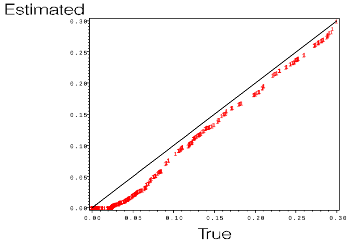


Figure 1e. Estimates vs Truth, File C
Cumulative Distribution of Matches
Large base, Independent EM, non−1−1

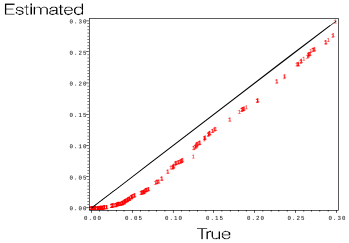


Figure 1b. Estimates vs Truth, File A
Cumulative Distribution of Nonmatches
Large base, Independent EM, non−1−1

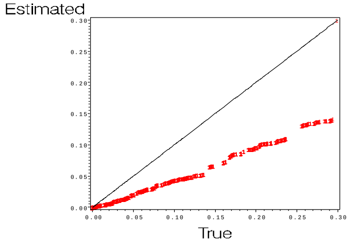


Figure 1d. Estimates vs Truth, File B
Cumulative Distribution of Nonmatches
Large base, Independent EM, non−1−1

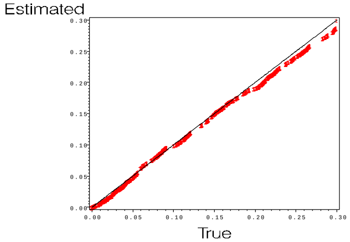


Figure 1f. Estimates vs Truth, File C
Cumulative Distribution of Nonmatches
Large base, Independent EM, non−1−1

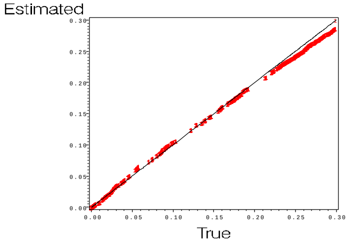


Figure 2a. Estimates vs Truth, File A
Cumulative Matches, Lambda=0.99
Small Sample, Independent EM, non−1−1

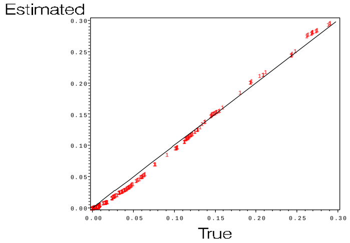


Figure 2c. Estimates vs Truth, File B
Cumulative Matches, Lambda=0.99
Small Sample, Independent EM, non−1−1

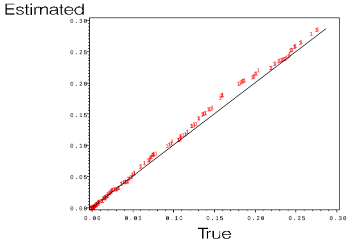


Figure 2e. Estimates vs Truth, File C
Cumulative Distribution of Matches, Lambda=0.99
Small Sample, Independent EM, non−1−1

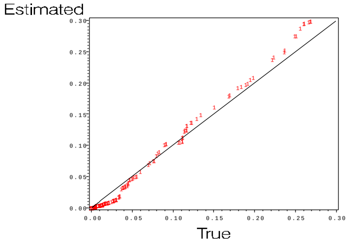


Figure 2b. Estimates vs Truth, File A
Cumulative Nonmatches, Lambda=0.99
Small Sample, Independent EM, non−1−1

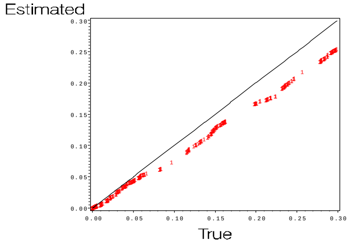


Figure 2d. Estimates vs Truth, File B
Cumulative Nonmatches, Lambda=0.99
Small Sample, Independent EM, non−1−1

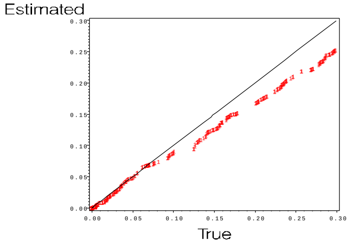
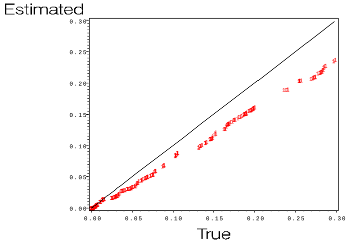


Figure 2f. Estimates vs Truth, File C
Cumulative Nonmatches, Lambda=0.99
Small Sample, Independent EM, non−1−1


Estimated
True