IMPROVED DECISION RULES IN THE FELLEGI-SUNTER MODEL OF RECORD LINKAGE

William E. Winkler*
Stat. Research Division, Room 3000-4, U.S. Bureau of the Census, Washington, DC 20233

ABSTRACT

Many applications of the Fellegi-Sunter model use simplifying assumptions and ad hoc modifications to improve matching efficacy. Because of model misspecification, distinctive approaches developed in one application typically cannot be used in other applications and do not always make use of advances in statistical and computational theory. This paper gives an Expectation-Maximization (EMH) algorithm that constrains the estimates to a convex subregion of the parameter space. The EMH algorithm provides probability estimates that yield better decision rules than unconstrained estimates. The algorithm is related to the Multi-Cycle Expectation-Conditional Maximization algorithm (Meng and Rubin 1993) and is deduced via results of Haberman (1977) that hold for large classes of loglinear models.

Key Words: MCECM Algorithm, Latent Class, Computer Matching, Error Rate

This paper provides a theory for obtaining constrained maximum likelihood estimates for latent-class, loglinear models on finite state spaces. The work is related to Expectation-Maximization (EM) algorithms by Meng and Rubin (1993) for obtaining unconstrained maximum likelihood estimates. Meng and Rubin generalized the original ideas of Dempster, Laird, and Rubin (1977), hereafter denoted by DLR. The new class of algorithms, denoted by EMH, apply results of Haberman (1977) to constrain estimates to appropriate subregions of the parameter space and assure that the likelihood of successive parameter estimates is nondecreasing. In a variety of cases including the one of this paper, the method of constraining estimates can be expressed in closed form. Thus, constraining estimates need not necessitate iterative fitting methods such as Newton-Raphson or grid search.

With many latent class models, computation and modelling is greatly simplified because observed variables are assumed to be independent conditional on unobserved classification variables (e.g., Titterington, Makov, and Smith 1988) or such independence can reasonably be assumed hold (Rubin and Stern 1993). With record linkage problems, conditional independence does not hold (Smith and Newcombe 1975, Thibaudeau 1993). If latent class models have a large, say ten, number of observed variables, modelling the correct set of interactions is considerably more difficult than it is with general loglinear models where is is known to be difficult (e.g., Bishop, Fienberg, and Holland 1975). Conventional statistics such as chi-square do not yield accurate indications of the fit of estimates to the truth (Winkler 1989, 1992, Rubin and Stern 1993).

Instead of modelling the precise set of interactions, it may be suitable to include an easily specified, say all 3-way, set of interactions and restrict the solutions to a subregion of the parameter space based on prior knowledge. If such constraints are appropriate, then the parameter estimates and decision rules may be nearly as good as those obtained through detailed modelling of specific sets of interactions. The constraints need to be easily specified and must be sufficiently weak that they provide sensible restriction in a variety of similar situations. Also, 3-way interactions and suitable parameter-space restrictions should yield reasonable approximations to true models with interactions higher than 3-way.

The main example involves a record linkage problem with files having known matching status. Computation is straightforward because each of the successive components of the Maximization step are in closed form and the restriction to a subregion of the parameter space is also in closed form using simple constraints. To motivate the concepts, basic parameter-estimation in record linkage and the successive types of EM-type estimation procedures that have so far been applied are described.

Fellegi and Sunter (1969) gave a formal model for record linkage that involves optimal decision rules that divide a product space A×B of pairs of records from two files A and B into matches and nonmatches, denoted by M and U, respectively. The main issue is the accuracy of estimates of probability distributions used in a crucial likelihood ratio. When estimates are sufficiently accurate, decision rules are (nearly) optimal. The optimality is in the sense that, for fixed bounds of the proportions of false matches and false nonmatches, the size of the set of pairs on which no decision is made is minimized.

Fellegi and Sunter (1969, pp. 1194-1197) considered the following probability decomposition

$$P(pat) = P(M)\, P(pat|M) + P(U)\, P(pat|U), \qquad (1.1)$$

where *pat* represents an agreement pattern on characteristics such as surname, house number, and phone. They observed that, in the case for which *pat* represents the eight patterns of simple agreement/disagreement on three fields and the agreements are conditionally independent given M and U, (1.1) represents seven equations and seven unknowns that can be solved directly. In general situations, the set of equations (1.1) can be solved by least squares, methods of moments, or maximum likelihood. Because methods of moments do not yield solutions that are as pleasing as those via maximum likelihood (Titterington, Smith, and Makov 1988), and least squares has shown numerical instability in record linkage situations (Jaro 1989), Expectation-Maximation (EM) algorithms (e.g., DLR) are used to get maximum likelihood estimates.

The second section of this paper presents the EMH algorithm. The third section gives more background on EM-type algorithms and the empirical data. The results in the fourth section compare estimates computed via the EMH algorithm with estimates computed via prior methods. In the fifth section, the limitations of current EM-type methods involving latent class models are described. The final section consists of a summary and conclusions.

2. EMH ALGORITHM

This section contains background on how EM-type procedures can be applied to latent class models. While the existing EM-type procedures are intended for unconstrained maximization on the parameter space Ω (DLR, Wu 1983, Meng and Rubin 1993), the EMH algorithm is intended to constrain solutions to a closed, convex subregion R of Ω. The key idea needed for the EMH algorithm is the following inequality due to Haberman.

Theorem. (Haberman 1975, 1977). Let the parameters being estimated via an EM-type procedure be products of multinomial or Poisson distributions. If ϕ_p and ϕ_{p+1} are successive estimates, then for all, $0 \leq \alpha \leq 1$, the log-likelihood L satisfies

$$L(\phi_p) \leq L(\alpha\ \phi_p + (1-\alpha)\ \phi_{p+1}). \qquad (2.1)$$

Inequality (2.1) states that all parameters on the line segment between ϕ_p and ϕ_{p+1} yield nondecreasing likelihood. If ϕ_p lies in the interior of a convex subregion R of the parameter space, then it is possible to obtain an α such that $\alpha\ \phi_p + (1-\alpha)\ \phi_{p+1}$ lies on the boundary of R. If the constraints defining R are simple, then such an α can be represented as a closed form solution of an equation; otherwise, such an α may have to be obtained via an iterative procedure such as Newton-Raphson. Inequality (2.1) does not hold for general EM-type procedures. In the following, the restraint functions $\{g_i, i = 1, ..., S\}$ can be assumed to be the same as those given by Meng and Rubin (1993, also 1991, pp. 245-246). While additional restraint functions may determine the closed, convex subregion R of the parameter space Ω, they are not explicitly needed in the statement of the algorithm. Replacing the missing data with expected values is referred to as *completing* or *filling-in* the data (DLR).

EMH Algorithm for loglinear models constrained to a closed, convex subregion R of parameter space Ω.

1. Beginning with an initial set of parameters ϕ_0 in R, complete data with expected values using ϕ_0.
2. Using the restraints imposed by the completed data and g_1, find the maximum likelihood estimate ϕ_1 in Ω. If $\phi_1 \notin R$, find the α so $\alpha\ \phi_0 + (1-\alpha)\ \phi_1$ is on the boundary of R and use $\alpha\ \phi_0 + (1-\alpha)\ \phi_1$ as the estimate. If $\phi_1 \in R$, use it as the estimate. Complete the data according to the new estimate ϕ_1.
3. Using the restraints imposed by the completed data and g_2, find the maximum likeliood estimate ϕ_2 in Ω. If $\phi_2 \in R$, use it as the estimate. If ϕ_1 is on the boundary of R and $\phi_2 \notin R$, use ϕ_1 as the estimate of ϕ_2. If ϕ_1 is in the interior of R and $\phi_2 \notin R$, find the α so $\alpha\ \phi_1 + (1-\alpha)\ \phi_2$ is on the boundary of R and use $\alpha\ \phi_1 + (1-\alpha)\ \phi_2$ as the estimate. Complete the data according to the new estimate ϕ_2.
4. Continue constrained E and M steps by successively cycling through all restraints in the manner of Step 3.

By the theorem, the EMH algorithm yields nondecreasing likelihood. If the restraint functions g_i, i = 1, 2, ..., S, are the usual constraints associated with iterative proportional fitting, then each conditional maximization is in closed form. If the constraints defining the subregion R are simple, then each of the αs that pull successive estimates back to the boundary of R are also in closed form. The initial estimate ϕ_0 must be in R. If there is no restriction to R, then the EMH algorithm corresponds to the MCECM algorithm of Meng and Rubin (1993). Rather than find a maximum at each CM-step, it is sufficient to find ϕ_{p+1} different from ϕ_p so that the likelihood increases. Computer code should monitor that the likelihood strictly increases on some of the CM-steps.

3. RECORD LINKAGE BACKGROUND

This section contains background on the previous applications of the EM to give insight into why the new methods were developed. The first subsection summarizes earlier EM-applications to record linkage and the second describes the empirical data.

3.1. Previous Applications of EM

The main reason that existing parameter-estimation procedures fail to yield optimal decision rules is that the conditional independence assumption is not valid. Thibaudeau (1993) observed that, when files contain name and address information, strong dependencies between agreements on fields such as surname, house number, street name, and phone number occur on the set on nonmatches U. Winkler (1992) showed that, instead of A×B naturally dividing into the desired two classes M and U, A×B can be naturally partitioned into three classes: C_1- matches agreeing on name and address, C_2- nonmatches agreeing on address, and C_3- nonmatches not agreeing on address. In the decision rules classes C_2 and C_3 are combined into U. The primary reason the 2-class EM procedure fails is that it divides the set of pairs into those agreeing on address and those not. If the 3-class EM algorithm is applied under the independence assumption, reasonable decision rules are obtained when matching decision thresholds are obtained manually (Winkler 1992). Error rates, however, cannot be estimated accurately. An alternative error-rate estimation procedure (Belin and Rubin 1993, Rubin and Belin 1991) can yield accurate estimated false match rates (Winkler and Thibaudeau 1991) in some cases but is not applicable to the situations of this paper.

Winkler (1992) applied 3-class EM algorithms under models in which all 3-way interactions were allowed. While the the interactions gave dramatically lower chi-square values and the overall fits as given by the estimated cumulative probability distributions appeared acceptable, both probability estimates for individual ageement patterns and the estimated proportion of pairs in class C_1 could differ substantially from the true values. The EM-type procedures arbitrarily classify sets of pairs into classes according to the variables that are agreeing in the patterns and according to interactions being fit. When a set of matching variables contains many associated with addresses, general EM-type procedures can yield probability estimates and resultant decision rules that give primary weight to address information and secondary weight to name and demographic information. When all 3-way interactions are fit, spurious agreements of nonessential variables may be given too much weight.

To better make use of prior information, specific sets of interactions can be modelled using new algorithms first applied by Armstrong (1992). The difficulty with using specific interactions is that there are far fewer available degrees of freedom (dofs) with latent class models than with ordinary loglinear models. For instance, with ten variables there are insufficient dofs to model all 4-way interactions; with eight variables, insufficient for all 3-way. If use of specific sets of interactions are not sufficient, then convex constraints can be used to predispose solutions to convex subregions that are more likely to yield accurate estimates.

3.2. Data Used in Results

The pairs are taken from two files having known matching status and 12,000 and 15,000 records, respectively. Only 116,305 pairs agreeing on a geographic identifier and the first character of the surname are used. There are 9800 matches. The analysis evaluates nearly all matches because less than 4 percent of the true matches disagree on the geographic identifier or on the first character of the surname. The matching fields that are: surname, first name, house number, street name, phone, age, relationship to head of household, marital status, sex, and race. To simplify computation all comparisons are considered agree/disagree. The ten data fields yield 1024 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 (15% blank) and phone (35% blank). Name and address data are never missing.

4. RESULTS USING EM-DERIVED PROBABILITIES

This section presents results from fitting using under five models: (1) independent, 3-class EM, (2) dependent, 3-class EMH with all 3-way interactions of variables, (3) dependent, 3-class EMH with a selected subset of interactions, and (4) dependent, 3-class EMH with all 3-way interactions and selected convex constraints, and (5) dependent, 3-class EMH with a selected subset of interactions and selected convex constraints.

When the number of interactions are increased, chi-square values will typically decrease. The selected interaction patterns (Table 4.1) are chosen as a compromise that allows a number of degrees of freedom so that the statistics can be tested. The first set of patterns were chosen with knowledge of some of the true statuses. The set of patterns is described as ja because combinatorial algorithms due to Armstrong (1992) are used in the fitting. The second set consisting of all 3-way interactions was chosen as a general exploratory tool that left some degrees of freedom for testing. Because the interaction patterns do not necessarily yield estimated probabilities that give good decision rules, the convex constraints of Table 4.2 are used to predispose estimates into subsets of the parameter space. The second set of convex constraints use twice the true underlying probabilities as an upper bound on the estimated probabilities. Some of

Table 4.1. Interactions Used in Fitting Hierarcharical Models

ja	1. last, first, hsnm, stnm, phone, age, rel, marit
	2. first, hsnm, phone, sex
	3. last, race
3-way	1. all 3-way

the convex constraints will lead to solutions that are on the boundary imposed by the constraints. Others will initially constrain the solutions to certain subsets but final limiting solutions will not hit the boundary.

Table 4.2. Convex Constraints Use in Fitting

ja
14 complicated restraints

3-way
$P(\{\text{agree last, disagree first}\} \cap C_1) \leq 0.0070$
$P(\{\text{disagree last, disag first}\} \cap C_1) \leq 0.0014$

The statistics associated with the fits of the different models yield somewhat contradictory information (Table 4.3). With the exception of the chi-squares associated with fits of the models that include all 3-way interactions, all chi-square values are much too high. The loglikelihood associated with fitting using the first set of interactions (denoted by ja) is closer to the theoretical maximum of -4.1071 than the log-likelihood from all 3-way interactions. The Z-statistic is the standard normal approximation to the chi-square statistic and is given as a reference. The reason that the Z-values associated with the ja models are higher than those of the 3-way models is that many more interactions are fit in the ja models and the Z-values involve fewer degrees of freedom. The P_1-value associated with Class C_1 is the estimated proportion of pairs that are matches M. The P_1-value 0.0878 associated with the first set of interactions (denoted by ja) is the second closest to the true P_1 value of 0.0869. The detailed constraints in ja, convex were chosen to force the estimated proportion P_1 close to true proportion.

Table 4.3. Summary Statistics Associated with Various Models

	log-like	chi-sq	Z	P_1
independent				
	-4.2206	26,383	570.4	.0910
ja				
	-4.1084	294	8.7	.0878
ja, convex				
	-4.1086	340	11.4	.0869
3-way				
	-4.1088	375	-2.8	.1015
3-way, convex				
	-4.1100	660	5.2	.0886

Plots of the cumulative probability distributions of the five models versus the truth (given by the 45 degree line) are presented in Figures 1-5 for matches; in Figures 6-10 for nonmatches. For matches, only the regions in which the error proportions (false nonmatch rates) are less than 0.30 are shown. With the exception of the independence model (Figures 1 and 6) which deviates substantially from the truth, all plots show reasonable fits to the true distributions. While the curves associated with the chosen subset of interactions (ja, Figures 4 and 9) appear acceptable, they mask the fact that some probability estimates at individual points deviate substantially from the truth. The final two models that involve convex constraints give better fits at individual points (and thus in each subrange of the distribution) than the others.

5. DISCUSSION

The results show that conventional chi-square statistics describing the fits give no valid indication of the quality of the estimated probabilities when they are used in decision rules. In decision rules, matches need to be distinguished from nonmatches. Often there is no clear demarcation between Class C_1, matches within the same household, and Class C_2, nonmatches within the same household. For instance, in one set of pairs, husband-wife pairs in which age agrees and sex agrees due to miskeying are placed in Class C_1; in other sets such pairs may be placed in Class C_2. This section describes limitations on the general applicability of convex constraints and the extension of the EMH algorithm to general statistics.

5.1. Basic Limitations on Use of Convex Constraints

With other similar data bases having the same matching variables, representing similar types of geographic characteristics, and for which true matching status was known, the following improving relative accuracies of estimated probabilities were observed

independent < all 3-way < ja selected interactions
< all 3-way + convex

where the ja selected interactions are the same as those given in Table 4.1 and the convex constraints are the same as those given in Table 4.2. Generally, the last two models had about the same accuracy and were much better than the first two. In one set of files in which some of the demographic variables had substantially higher typographical variation than in other files, the last model yielded more accurate estimated probabilities.

On an absolute basis, the estimated probabilities under the last two models were not as accurate as those given Figures 3-5 and 8-10 but were still reasonable. The estimated cumulative probabilities given a nonmatch were typically much more accurate than the corresponding cumulative probabilities given a match.

No set of convex constraints or set of interactions has been found that consistently yield highly accurate estimates. This is is due to the fact that the underlying true probabilities vary significantly from data set to data set. With ten matching variables, some true conditional probabilities associated with individual data patterns vary by an order of magnitude. The estimates associated with the ja selected interaction model exhibited more variation than estimates under the all 3-way, convex model. Severe or unusual typographical variation in only 0.1 percent of the pairs (associated with Classes C_1 and C_2) were sufficient for significant changes in estimates.

5.2. EMH Algorithm for General Statistics

The idea of conditional maximization in which parameters are constrained to convex subregions of the parameter space can be extended from latent class models to general statistics. The basic idea is still the one originally emphasized by Meng and Rubin. It is reduce dimensionality to make computation associated with maximizations faster or more stable. In general, it is possible to search for the α_0 so that $\alpha_0\ \phi_p + (1-\alpha_0)\ \phi_{p+1}$ lies in the subregion R imposed by a set of constraints and (approximately) maximizes the likelihood over all α, $0 \leq \alpha \leq 1$. As it is only necessary to increase the likelihood, it may often be possible to find a simple algorithm to obtain an $\alpha_0 > 0$ that yields increasing likelihood. In general, such an α_0 can be found via one-dimensional Newton-Raphson or grid-search methods.

6. SUMMARY AND CONCLUSIONS

This paper describes general theory and algorithms for fitting loglinear models for latent classes. The algorithms do not require an independence assumption on the estimated conditional probabilities associated with different latent classes, are related to ideas of Haberman (1977) and Meng and Rubin (1993), and allow convex constraints to be imposed on the estimated probabilities. Because conventional chi-square and other statistics describing the fits do not give good indication of the accuracy of the estimated probabilities, the convex constraints can be used to predispose solutions to subregions of the parameter space that are consistent with prior knowledge.

*This paper reflects views of the author and not necessarily those of the Bureau of the Census. A longer version of this paper is available from the author.

REFERENCES

Armstrong, J. A., (1992), "Error Rate Estimation for Record Linkage: Some Recent Developments," in Proceedings of the Workshop on Statistical Issues in Public Policy Analysis, Carleton University.

Belin, T. R. and Rubin, D. R. (1993), "A Method for Calibrating False-Match Rates in Record Linkage", technical report, Dept. of Biomathematics, UCLA.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W., (1975), Discrete Multivariate Analysis, Cambridge, MA: MIT Press.

Dempster, A. P., Laird, N. M., and Rubin, D. B., (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," J. of the Royal Stat. Soc. B, 39, 1-38.

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

Haberman, S. J., (1975), "Iterative Scaling for Log-Linear Model for Frequency Tables Derived by Indirect Observation," Proceedings of the Section on Statistical, Computing, American Statistical Association, 45-50.

Haberman, S. J., (1977), "Product Models for Frequency Tables involving Indirect Observation," Annals of Statistics, 5, 1124-1147.

Jaro, M. A., (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," Journal of the American Statistical Association, 89, 414-420.

Meng, X. L., and Rubin, D. B., (1993), "Maximum Likelihood Via the ECM Algorithm: A General Framework," Biometrika, to appear.

Rubin, D. B., and Belin, T. R., (1991), "Recent Developments in Calibrating Error Rates for Computer Matching," Proceedings of the 1991 Census Annual Research Conference, 657-668.

Rubin, D. B., and Stern, H. S. (1993), "Testing in Latent Class Models Using a Posterior Predictive Check Distribution," Analysis of Latent Variables in Development Research, (A. von Eye and C. Clogg (eds)).

Thibaudeau, Y., (1993), "The Discrimination Power of Dependency Structures in Record Linkage," Survey Methodology, 19, 31-38.

Titterington, D. M., Smith, A. F. M., Makov, U. E., (1988), Statistical Analysis of Finite Mixture Distributions, New York: J. Wiley.

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

Winkler, W. E., (1992), "Comparative Analysis of Record Linkage Decision Rules," Proc. of the Section on Surv. Res. Methods, American Statistical Assoc., 829-834.

Winkler, W. E., and Thibeaudeau, Y. (1991), "An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Decennial Census," Statistical Research Division Report, U.S. Bureau of the Census.

Wu, C. F. J., (1983), "On the convergence properties of the EM algorithm," Annals of Statististics, 11, 95-103.

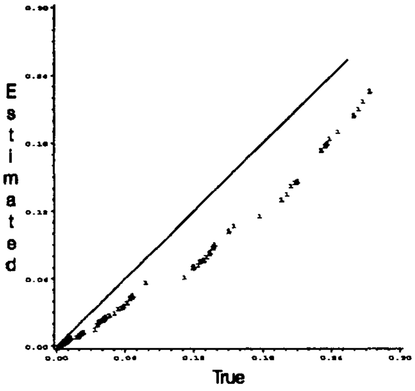

Figure 1. Estimates vs Truth
Cumulative Distribution of Matches
3-Class, Independent EM

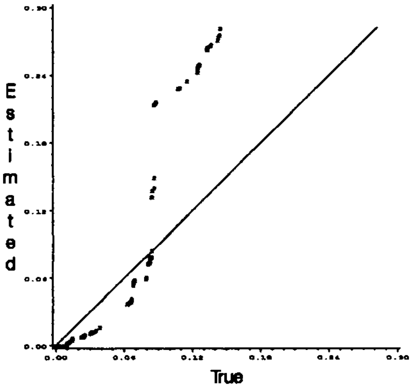

Figure 2. Estimates vs Truth
Cumulative Distribution of Matches
3-Class, 3-way Interaction EM

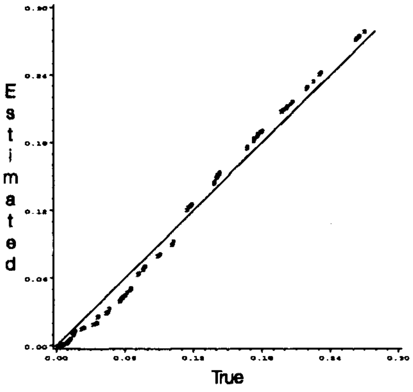

Figure 3. Estimates vs Truth
Cumulative Distribution of Matches
3-Class, 3-way Interaction EM, Convex

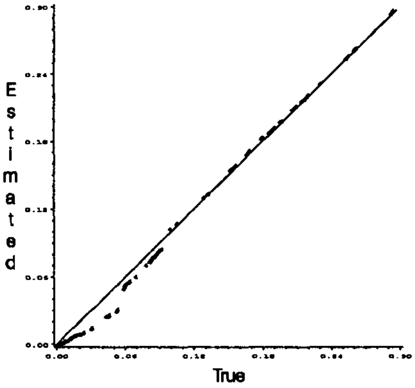

Figure 4. Estimates vs Truth
Cumulative Distribution of Matches
3-Class, ja Interaction EM

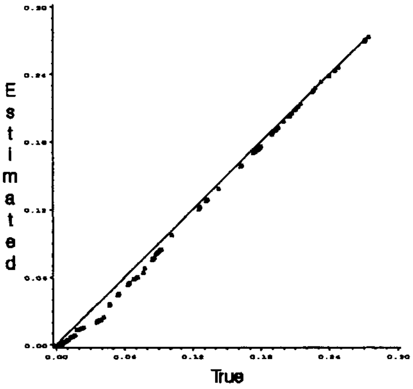

Figure 5. Estimates vs Truth
Cumulative Distribution of Matches
3-Class, ja Interaction EM, Convex

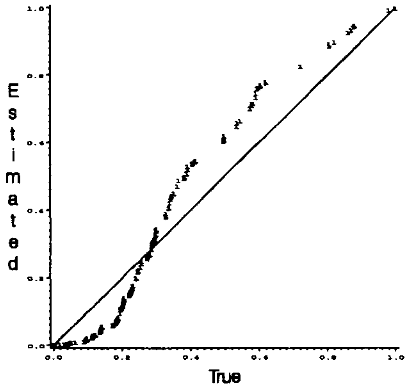

Figure 6. Estimates vs Truth
Cumulative Distribution of Nonmatches
3-Class, Independent EM

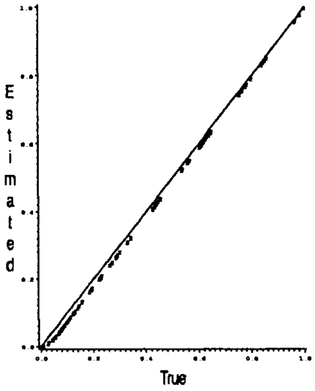

Figure 7. Estimates vs Truth
Cumulative Distribution of Nonmatches
3-Class, 3-way Interaction EM

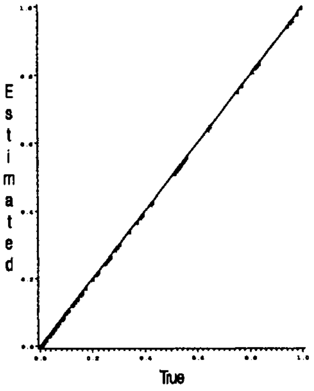

Figure 8. Estimates vs Truth
Cumulative Distribution of Nonmatches
3-Class, 3-way Interaction EM, Convex

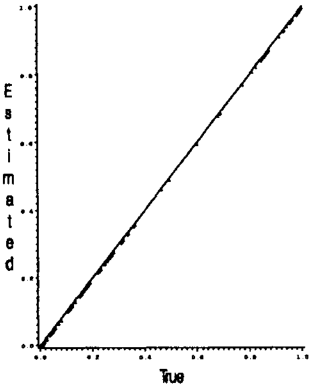

Figure 9. Estimates vs Truth
Cumulative Distribution of Nonmatches
3-Class, ja Interaction EM

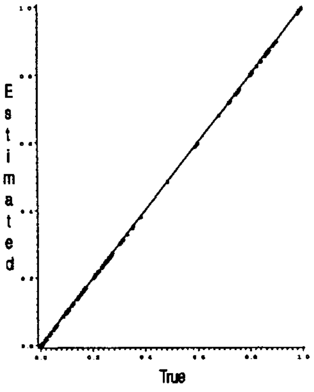

Figure 10. Estimates vs Truth
Cumulative Distribution of Nonmatches
3-Class, ja Interaction EM, Convex