USING THE EM ALGORITHM FOR WEIGHT COMPUTATION IN THE FELLEGI-SUNTER MODEL OF RECORD LINKAGE

William E. Winkler, Bureau of the Census *
Washington, D.C. 20233

KEY WORDS: Decision rule, error rate.

1. INTRODUCTION

The paper describes using the EM Algorithm (Dempster, Laird, and Rubin 1977, Haberman 1977, Wu 1983) for parameter estimation in the Fellegi-Sunter model of record linkage. The method is applicable for more general classes of distributions than those considered by Fellegi and Sunter (1969).

Let A X B be the product space of two sets A and B which is divided into matches (pairs representing the same entity) and nonmatches (pairs representing different entities). Linkage rules are those that divide A X B into links (designated matches), possible links (pairs for which we delay a decision), and nonlinks (designated nonmatches).

Under fixed bounds on the error rates, Fellegi and Sunter (1969, hereafter denoted FS) provided a linkage rule that is optimal in the sense that it minimizes the set of possible links. The optimality is dependent on knowledge of certain probabilities that are used in a crucial likelihood ratio.

In applications, an independence assumption is made that allows estimation of the probabilities. The probabilities are referred to as matching parameters. If the independence assumption is not valid (Winkler 1985, 1987; Kelley 1986), then linkage rules based on the estimated probabilities may not be optimal.

The remainder of this paper contains a methodology for estimating parameters for general distributions. Section two is divided into four parts. The first part provides a summary of the FS Model of record linkage. The second describes the basic EM Algorithm when the underlying distributions are independent. Computation is particularly straight-forward because of the closed-form maximization step.

The third part describes the EM Algorithm for general distributions. The theoretical validity can be deduced from methods of Haberman (1977, 1979) or more generally Dykstra (1985a, b, 1987, 1988). The fourth part presents a procedure for deriving frequency-based weights when a weak independence assumption is met. The assumption involves only two parameters, is weaker than the assumption of FS (pp. 1207-1210), and can typically be shown to hold in practice.

The discussion in the third section comprises three components. The first describes convergence properties of the EM Algorithm. The second describes Dykstra's computational methods which generalize some iterative fitting methods than have typically been used (e.g., Haberman 1977, 1979). The third discusses the computation of frequency-based weights.

The final section is a summary.

2. MODEL AND COMPUTATIONAL PROCEDURES

2.1. Fellegi-Sunter Model

The FS Model uses an decision-theoretic approach embodying principles first used in practice by Newcombe (Newcombe et al. 1959). To give an overview, we describe the model in terms of ordered pairs in a product space. The presentation closely follows FS (pp. 1184-1187).

There are two populations A and B whose elements will be denoted by a and b. We assume that some elements are common to A and B. Consequently the set of ordered pairs

$$A \times B = \{(a,b) : a \in A,\ b \in B\}$$

is the union of two disjoint sets of matches

$$M = \{(a,b): a=b,\ a \in A,\ b \in B\}$$

and nonmatches

$$U = \{(a,b): a \neq b,\ a \in A,\ b \in B\}.$$

The records corresponding to A and B are denoted by $\alpha(a)$ and $\beta(b)$, respectively. The comparison vector τ associated with the records is defined by:

$$\tau[(\alpha(a),\beta(b))] \equiv \{\tau^1[(\alpha(a),\ \beta(b))],$$

$$\tau^2[\alpha(a),\beta(b))],\ldots,\tau^K[(\alpha(a),\beta(b))]\}.$$

Where confusion does not arise, the function τ on A X B will be denoted by $\tau(\alpha,\beta)$, $\tau(a,b)$, or τ. The set of all possible realization of τ is denoted by Γ.

The conditional probability of $\tau(a,b)$ if $(a,b) \in M$ is given by

$$m(\tau) \equiv P\{\tau[\alpha(a),\beta(b)]:(a,b)\in M\}$$

$$= \sum_{(a,b)\in m} P\{\tau[\alpha(a),\beta(b)]\} \cdot P[(a,b)|M].$$

Similarly we denote the conditional probability of τ if $(a,b) \in U$ by $u(\tau)$.

We observe a vector of information $\tau(a,b)$ associated with pair (a,b) and wish to designate a pair as a link (in set A_1), a possible link (in set A_2), or a nonlink (in set A_3). We let L denote a linkage rule that divides A X B into A_1, A_2, and A_3). We say that a Type I error has occurred if rule L places $m \in M$ in A_3,

$$P(A_3|M) = \sum_{\tau\in\Gamma} m(\tau)\ P(A_3|\tau),$$

and a Type II error if L places $u \in U$ in A_1,

$$P(A_1|U) = \sum_{\tau\in\Gamma} u(\tau)\ P(A_1|\tau).$$

FS define a linkage rule L_0 with associated sets A_1, A_2, and A_3 that is optimal in the following sense:

THEOREM (Fellegi and Sunter 1969). Let L' be

a linkage rule with associated sets A_1', A_2', and A_3' such that $P(A_3'|M) = P(A_3|M)$ and $P(A_1'|U) = P(A_1|U)$. Then $P(A_2|U) \leq P(A_2'|U)$ and $P(A_2|M) \leq P(A_2'|M)$.

In other words, if L' is any competitor of L_0 having the same Type I and Type II error rates (which are both conditional probabilities), then the conditional probabilities (either on set U or M) of not making a decision under rule L' is always greater than under L_0.

The FS linkage rule is actually optimal with respect to any set Q of ordered pairs in A X B if we define error probabilities P_Q and a linkage rule L_Q conditional on Q. Thus, it may be possible to define subsets of A X B on which we make use of differing amounts and types of available information.

In application, we consider the following likelihood ratio

$$R \equiv R[\tau(a,b)] = m(\tau)/u(\tau). \qquad (2.1)$$

If the numerator is positive and the denominator is zero in (2.1), we assign a fixed very large number to the ratio. The FS linkage rule takes the form:

If R > UPPER, then denote (a,b) as a link.
If LOWER ≤ R ≤ UPPER, then denote (a,b) as a possible link. (2.2)
If R < LOWER, then denote (a,b) as a nonlink.

The cutoffs LOWER and UPPER are determined by the desired error rate bounds.

2.2 EM Algorithm for Independent Distributions

Applying the FS model involves determining estimates of the conditional probabilities $m(\tau)$ and $u(\tau)$. To obtain maximum likelihood estimates we use the EM Algorithm.

For record pairs r_j, j = 1, 2,...,N, from Q, index the comparison vectors τ_j^i as follows:

τ_j^i = 1 if field i agrees for record pair r_j.
= 0 if field i disagrees for record pair r_j.

The elements in $Q = (Q \cap M) \cup (A \cap U)$ are distributed according to a finite mixture with the unknown parameters $\Phi = (m,u,p)$ where p is the proportion of matched pairs in Q. Let x be the complete data vector $g = \langle \tau_j, g_j \rangle$ where

$g_j = (1,0)$ if $r_j \in M \cap Q$ and

$g_j = (0,1)$ if $r_j \in U \cap Q$.

Then the complete data log-likelihood (Dempster, Laird, and Rubin 1977, pp. 15-16) is given by

$$\ln f(x|\phi) = \sum_{j=1}^{N} g_j \cdot \langle \ln P(\tau_j \mid M \cap Q), \ln P(\tau_j | U \cap Q) \rangle + \sum_{j=1}^{N} g_j \cdot \langle \ln p, \ln(1-p) \rangle.$$

Fitting using the EM Algorithm will be performed under the following assumption:
There exist vector constants $m \equiv (m_1, m_2, \ldots, m_K)$ and $u \equiv (u_1, u_2, \ldots, u_K)$ such that, for all $\tau \in \Gamma$,

$$P(\tau \mid M \cap Q) = \prod_{i=1}^{K} m_i^{\tau^i} (1-m_i)^{(1-\tau^i)}$$

and (2.3)

$$P(\tau \mid U \cap Q) = \prod_{i=1}^{K} u_i^{\tau^i} (1-u_i)^{(1-\tau^i)}.$$

Probabilities m_i and u_i, i = 1,2,...,K, are constant for all representations τ of pairs in Q. To avoid trivialities, we assume that $(0<m_i)$, $u_i<1$, i = 1,2,...,K.

We begin the EM Algorithm with estimates of the unknown parameter $\langle \hat{m}, \hat{u}, \hat{p} \rangle$. For the E-step under (2.3), replace g_j with $\langle P(M \cap Q|\tau_j), P(U \cap Q|\tau_j) \rangle$ where

$$\hat{P}(M \cap Q|\tau_j) \equiv \frac{\hat{p} \prod_{i=1}^{K} \hat{m}_i^{\tau_j^i} (1-m_i)^{(1-\tau_j^i)}}{D}$$

and $\hat{P}(U \cap Q|\tau_j) = 1 - \hat{P}(M \cap Q|\tau_j)$,

where (2.4)

$$D = \hat{p} \prod_{i=1}^{K} \hat{m}_i^{\tau_j^i} (1-m_i)^{(1-\tau_j)} + (1-\hat{p}) \prod_{i=1}^{K} \hat{u}_i^{\tau_j^i} (1-\hat{u}_i)^{(1-\tau_j^i)}.$$

For the M step, the complete data log-likelihood can be separated into three maximization problems. Setting the partial derivatives equal to zero and solving for $\hat{m}_i$, i = 1, 2,...,K, yields:

$$\hat{m}_i = \frac{\sum_{j=1}^{N} \hat{P}(M \cap Q \mid \tau_j) \cdot \tau_j^i}{\sum_{j=1}^{N} \hat{P}(M \cap Q \mid \tau_j)} \qquad (2.5)$$

Estimates $\hat{u}_i$, i = 1, 2,...,K, are derived in a similar manner. The matrix of second partial derivatives can be shown to be negative-definite. The estimate of the proportion of matched pairs is given by

$$\hat{p} = \frac{\sum_{j=1}^{N} \hat{P}(M \cap Q \mid \tau^j)}{N}.$$

2.3. EM Algorithm for General Distributions

For general estimation of $m(\tau)$ and $u(\tau)$ we apply a generalized version of Theorem 4 of Haberman (1977). The generalization involves assuming that the proportion p of matches is bounded above and that the matrix of observed population frequencies can contain zeros.

The E-step takes the form

$$\hat{P}(M \cap Q|\tau_j) = \frac{\hat{p}\ P(\tau_j|M \cap Q)}{\hat{p}\ \hat{P}(\tau_j|M \cap Q) + (1-\hat{p})\ \hat{P}(\tau_j|U \cap Q)}$$

and (2.6)

$$\hat{P}(U \cap Q|\tau_j) = 1 - \hat{P}(M \cap Q|\tau_j).$$

The M-step involves finding maximum likelihood estimates $\hat{P}(\tau_j|M \cap Q)$, $\hat{P}(\tau_j|U \cap Q)$, and $\hat{p}$. Generally, the estimates must be found using iterative fitting procedures (e.g. Haberman 1976, 1977, 1979). Fitting satisfies the restraint that the margins are the fixed values determined by the observed frequency patterns and the p is constrained to be less than a fixed upper bound. The number of interaction terms used in fitting determine the number of dependent relationships.

2.4. Extension to Frequency-Based Weights

This section considers a procedure for extending simple agreement/disagreement weights to weights that account for frequency. We call such a procedure a dispersion. When the more stringent assumptions of FS (pp. 1207-1210) are satisfied our dispersion procedure agrees with theirs. If the agreement/disagreement weights found via the EM Algorithm coincide with the agreement/disagreement weights found via the FS procedures, then the frequency-based weights also coincide.

Frequency-based weights are useful because they can account for the fact that a specific surname pair such as (Zabrinksy, Zabrinsky) occurs less often than a surname pair such as (Smith, Smith).

We need some background material before presenting the computational procedures for frequency-based weights.

We observe that if, for some i and k,

$$m_i = P(\tau^k = 1 \mid M \cap Q)$$

and (2.7)

$$u_i = P(\tau^k = 1 \mid U \cap Q),$$

then the kth comparison is independent of the other K-1 comparisons.

The right hand sides of (2.7) are just the appropriate marginal inclusion probabilities.

Note that m_i and u_i, i = 1, 2,...,K of this paper generally differ from the m_1, m_2, m_3, u_1, u_2, and u_3 in FS (pp. 1194-1195, 1207-1210).

We define a random variable $\hat{\tau}^k$ by $\hat{\tau}^k = \mu_j^k$ if the kth comparison pair takes value μ_j^k where μ_j^k, j = 1,..., L^k, is an enumeration of the specific values of the kth comparison. We make two assumptions:

A1. Agreement/disagreement in the kth comparison is independent of the other K-1 comparisons.

A2. There exists a comparison k' such that the specific realizations of $\hat{\tau}^k$ are pairwise independent of agreement/disagreement in the k'th comparison.

If we consider one comparison, say of agreement/disagreement in surname, then we can perform EM fitting under a restricted version of (2.3) by specifying that one of the (m_i, u_i) must converge to the marginal probabilities (as in (2.7)) associated with surname. We can, thus, always find a comparison satisfying assumption A1 for the restricted class of distributions.

Assumption A2 is a weaker form of independence assumption than the one considered by FS (p. 1208). It allows dispersion of the agreement/disagreement weight obtained under assumption A1 to frequency-based weights.

In a manner similar to the dispersion of FS (pp. 1207-1210), we define

$$N_k(\mu_i^k) = P(\hat{\tau}^k = \mu_i^k, \tau^{k'} = 1),$$

$$V_k(\mu_i^k) = P(\hat{\tau}^k = \mu_i^k),$$

c = # pairs in Q, and

N = # pairs in $M \cap Q$.

Then, for i = 1, 2,...,L^k,

$$c \cdot N_k(\mu_i^k) = N \cdot P(\hat{\tau}^k = \mu_i^k \mid M \cap Q) \cdot P(\tau^{k'} = 1 \mid M \cap Q) + (c - N) \cdot P(\hat{\tau}^k = \mu_i^k \mid U \cap Q) \cdot P(\tau^{k'} = 1 \mid U \cap Q) \quad (2.8)$$

and

$$c \cdot V_k(\mu_i^k) = N \cdot P(\hat{\tau}^k = \mu_i^k \mid M \cap Q) + (c-N) \cdot P(\hat{\tau}^k = \mu_i^k \mid U \cap Q). \quad (2.9)$$

In (2.8) and (2.9) c, $N_k(\mu_i^k)$, $V_k(\mu_i^k)$, i=1,2,...,L^k, can be computed directly because they are based on observed file characteristics. The marginal probabilities $P(\tau^k=1 \mid M \cap Q)$ and $P(\tau^k=1 \mid U \cap Q)$ and the number of matches N in $M \cap Q$ can be computed using the estimated parameters of (2.4) that are obtained by the EM Algorithm. Equations (2.8) and (2.9), thus, consist of two equations to be solved for the two unknowns $P(\hat{\tau}^k = \mu_i^k \mid M \cap Q)$ and $P(\hat{\tau}^k = \mu_i^k \mid U \cap Q)$, i = 1, 2,...,$L^k$.

3. DISCUSSION

This section is divided into three parts. The first discusses the convergence properties of the EM Algorithm. The second describes a general computational methods due to Dykstra, Lemke, and Wollan. The third considers the extension to frequency-based weights.

3.1 Convergence Properties of EM Algorithm

This paper's application of the EM Algorithm most closely resembles the approach of Haberman (1977). Although Haberman's EM convergence proof was proved under more restrictive assumptions than the assumptions of this paper, it can be extended to deal with bounds on the proportion of matches. Haberman observed that the limiting value was dependent of the initial values of the parameters and, thus, not necessarily unique.

Wu (1983) noted that limiting values of the EM Algorithm are stationary points that can either be saddle points or local maxima. He made the conjecture that there is unlikely to be any general condition that assures convergence to a unique maximum.

Wu did observe, however, that if the likelihood is unimodal, if the estimated parameters have at most one stationary limiting point, and if a technical condition holds (which it does for the distributions of this paper), then the estimated parameters converge to the unique maximizer of the likelihood.

The implication is that, while the EM algorithm of this paper is of value in accounting for failures of the Conditional Independence Assumption, several starting points for the EM Algorithm should be used. The estimated parameters associated with the largest local maximum are the ones that are used.

If we can show that there is at most one stationary limiting point, then the parameter estimates will necessarily converge to it.

3.2. Dykstra's Computational Procedure

Dykstra and Lemke (1988) have shown the duality of maximum likelihood estimates and I-Projections under cone constraints. The cone constraints are more general than the constraints considered by Haberman (1977, 1979) and in this paper.

On the dual space, computation of individual I-Projections under affine constraints (Dykstra 1985a) is sometimes much easier than computation of the corresponding maximum likelihood estimates. Iteratively computing a limiting solution satisfying multiple restraints can be done using an algorithm of Dykstra and Wollan (1987). The theoretical validity of the algorithm follows from Dykstra's Iterative Fitting Procedure (1985b).

3.3. Extension to Frequency-Based Weights

Under the more stringent assumptions of FS (pp. 1207-1210) frequency-based weights computed using the techniques of this paper agree with those computed using Method II of FS. This follows because the dispersion method of this paper is identical to the dispersion method of FS and there can exist at most one local maximum of the likelihood.

The chief value of assumptions like Assumptions A1 and A2 of this paper is that they allow dispersal of agreement/disagreement weights with little increase in computation. Although the EM Algorithm might be extended to allow direct computation of frequency-based weights, such an extension will generally require enormous increases in computation.

4. SUMMARY

This paper describes using the EM Algorithm for estimating matching weights in the Fellegi-Sunter model of record linkage. The general theoretical and computational validity can be deduced using techniques of Haberman (1977, 1979) and of Dykstra (1985a, b, 1987, 1988). The procedure automatically incorporates a Bayesian adjustment for margins in the matrix of observed population frequencies.

5. ACKNOWLEDGEMENT

The author would like to thank Dr. R. Patrick Kelley, who first used the EM Algorithm in computing weights, for a number of comments.

REFERENCES

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

Dykstra, R. (1985a). "Computational Aspects of I-Projections," J. Statistical Computation and Simulation. 21 265-274.

Dykstra, R. (1985b). "An Iterative Procedure for Obtaining I-Projections onto the Intersection of Convex Sets," Ann. Prob. 13 975-984.

Dykstra, R. and Lemke, J. (1988). "Duality of I Projections and Maximum Likelihood Estimation for Log-Linear Models Under Cone Constraints," JASA 83 546-554.

Dykstra, R. and Wollan, P. (1987). "Finding I-Projections Subject to a Finite Set of Linear Inequality Constraints," Applied Statistics. 36 377-383.

Haberman, S. J. (1976), "Interative Scaling for Log-Linear Models for Frequency Tables Derived by Indirect Observation, ASA 1976 Proceedings of the Section on Statistical Computing, 45-50.

Haberman, S. J. (1977), "Product Models for Frequency Tables Involving Indirect Observation: Maximum Likelihood Equations," Ann. Stat. 5 1124-1147.

Haberman, S. (1979), Analysis of Qualitative Data, Academic Press. New York.

Kelley, R. P. (1986), "Robustness of the Census Bureau's Record Linkage System, ASA 1986 Proceedings of the Section on Survey Research Methods, 620-624.

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

Winkler, W. E. (1985), "Exact Matching Lists of Businesses: Blocking Subfield Identification, and Information Theory," ASA 1985 Proceedings of the Section on Survey Research Methods, 438-443.

Winkler, W. E. (1987), "Computational Aspects of Applying the Fellegi-Sunter Model of Record Linkage to Lists of Businesses," paper presented at the Symposium on Statistical Uses of Administrative Data. Ottawa, Ontario, Canada. November 1987.

Wu, C. F. J. (1983) "On the Convergence Properties of the EM Algorithm," Ann. Stat. 11 95-103.

* This paper reports the general results of research undertaken by the Census Bureau staff. The views expressed are attributable to the author and do not necessarily relect those of the Census Bureau.