A partial order approach to record linkage

D. H. Judson[*]

10/6/99; Short FCSM version: 9/10/01

Abstract

This paper applies the theory of semi-coherent structures (also known as monotonic boolean functions) to the problem of linking records across databases. We use the parameter estimates from a best fitting generalized linear model to derive the equivalent semi-coherent structure function, which we then call our best fitting record linkage rule. In this paper, we first describe the application area of record linkage, followed by a description of the Fellegi-Sunter model of record linkage. We then describe how the record linkage problem can be encompassed in partial order theory, and use the theory of semi-coherent structures to develop a parallel record linkage approach. We develop two methods for implementing the semi-coherent structure approach: 1) Estimate an inequality constrained generalized linear model and, using the parameter estimates, determine the semi-coherent structure that best fits the training data; and 2) Estimate a Bayesian generalized linear model and, using the posterior kernels, determine the semi-coherent structure that best fits the data. The best-fitting semi-coherent structure, derived from each approach, is our final estimated decision rule. We illustrate the approaches with analyses of a factorial experiment.

1. A Description of Record Linkage Problems

Imagine two databases, database A and database B, each of which contain information on individual persons. (We note for the record that it is not important

[*] Dean H. Judson (Dean.H.Judson@census.gov) is Mathematical Statistician and Group Leader, Administrative Records Evaluation and Linkage Group, Planning Research and Evaluation Division, U.S. Census Bureau. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review by the Census Bureau than its official publications. This report is released to inform interested parties and to encourage discussion. *Please note that this FCSM paper is a highly abbreviated version of a longer research paper available from the author. In particular, proofs, results, and references omitted in this paper are available in the longer version.*

that the object of the database is a person; it could be housing units or establishments or any other object of interest.) Suppose that each of these databases contains some information not contained in the other. If the databases refer to the same population, it would be of some analytic value to combine the separate information on each person into a single record. Presumably, we can then tabulate characteristics from the first database with characteristics from the second, and *ipso facto* we are able to make broader statements from the joint database than we could make from either separately.

Suppose further that each database does not contain an unique identifier—there is no "SSN" or comparable field in each database, but suppose that there are several fields which, *when used together*, might be sufficient to uniquely match cases. (Notably, this is the same problem as that of protecting the identity of an individual person in a microdata file release [11], only in reverse—we *want* to identify the person uniquely.)

The above is a simplified description of the fundamental "record linkage problem" ([9], [30]). For the purposes of this description, we have ignored some typical problems, for example, address standardization issues, street name conventions, differences in coding conventions across different databases, criteria for calling two names with typographical errors a match, and the like. These are not trivial matters; however, our emphasis in this paper is on the *algebraic structure* of the problem, not particular implementation issues.

2. The Fellegi-Sunter model of record linkage

The Fellegi-Sunter (which we shall abbreviate F-S on occasion) model is framed as an "hypothesis test" when two records are compared. The separate fields are compared on a field-by-field basis, and the F-S model uses information about the relative frequency of those fields to output one of three "decisions": The records are declared to be a positive match; the records are declared to be a positive nonmatch, or the decision "Don't know" is returned ([9]). The "don't know" region is then sent to presumably expensive clerical review and resolution. If we (arbitrarily) label positive match with 1, don't know with 2, and positive nonmatch with 3, then the F-S model is a function from the space $A \times B$ into the space $\{1, 2, 3\}$.

In the general F-S model, the problem is framed as an ordering of configurations by their "weight." For any collection of N individual fields $\overrightarrow{x}$, that configuration gets a ratio of match weight and nonmatch weight, $w\left(\overrightarrow{x}\right) = m(\overrightarrow{x})/u\left(\overrightarrow{x}\right)$. The configurations are presumed ordered by this ratio (ties broken arbitrarily), indexed in order from highest weight to lowest weight, and cutoffs are chosen defining a function $f(\overrightarrow{x})$ as:

$$f(\overrightarrow{x}) = \begin{cases} \text{Positive match} & \text{if } \ i \geq \text{upper cutoff} \\ \text{Positive nonmatch} & \text{if } i \leq \text{lower cutoff} \\ \text{Don't know} & \text{if upper cutoff} > i > \text{lower cutoff} \end{cases} .$$

(Here we are using F-S's corollary 1 to ignore randomized choices falling exactly on a boundary.)

The only remaining task is to define the weights w. In the standard F-S model, denoting the event "the records are in fact a match" by M and "the records are in fact a nonmatch" by $\tilde{}M$, these weights are defined as likelihood ratios, viz.,

$$w\left(\overrightarrow{x}\right) = \frac{P\left[\overrightarrow{x} \text{ configuration}|M\right]}{P\left[\overrightarrow{x} \text{ configuration}|\tilde{}M\right]} \tag{2.1}$$

The F-S paper ([9]) demonstrated that these weights are optimal in the sense that, for a fixed false match rate α and fixed false non-match rate μ, this decision rule, using these weights, minimize the clerical review region.

3. The implications (and limitations) of conditional independence

3.1. Conditional independence

The Fellegi-Sunter model of record linkage remains the standard today. Variations have been proposed on how weights should be calculated or used ([29]), but fundamentally the notion of an "hypothesis test" involving three regions, and a mapping from the space of field comparisons into the space of record linkage decisions, has not been modified since its original proposal.

Definition 3.1. *Suppose we have two or more databases with N fields on which we wish to perform matching. We will call our matching fields $x_1 \ldots\ x_N$, and define $x_i = 1$ if the two ith fields match, and $x_i = 0$ if the two ith fields do not match.*

Note that we have ignored the practical problem of exactly defining when two fields match. For the purposes of visualizing this system, one can consider a "match" to have occurred when the two fields are the same, i.e., FIRSTNAME="Joan" in database A and FIRSTNAME="Joan" in database B. In actual practice, one might prefer to call two fields matching under some less restrictive criterion. In practice, this is substantially more problematic that it might appear on first consideration. For the purposes of this theory, however, we are going to sidestep such practical matters.

In practice, however, an important additional assumption is often made: Conditional independence. We shall refer to models making the conditional independence assumption as CI models. This assumption allows the weights in the equation above to be factored into:$w\left(\overrightarrow{x}\right) = \frac{P[x_1=1|M]}{P[x_1=1|\tilde{}M]}\frac{P[x_2=1|M]}{P[x_2=1|\tilde{}M]}\cdots\frac{P[x_N=1|M]}{P[x_N=1|\tilde{}M]}$; when logarithms are taken, as is typically done in practice, this becomes a sum.

In order to identify the theoretical limits of the conditional independence assumption, it is necessary to define more terminology.

Definition 3.2. *We call a vector* $\overrightarrow{x}$ boolean *if it consists of N individual boolean variables, thus* $\overrightarrow{x} = (x_1, x_2, ...x_N)$ *. We call a function* $f : X \to Y$ boolean *if the domain of the function consists of vectors* $\overrightarrow{x} = (x_1, x_2, ...x_N)$ *, with* $x_i \in \{0, 1\}$ *, and the codomain of the function consists of a single boolean variable* $y \in \{0, 1\}$ *.*

Definition 3.3. *We call a boolean function* $f : X \to Y$ monotonic *if, for all* $\overrightarrow{x}$ *in the domain of f, and each* x_i*,* $f(\overrightarrow{x})|_{x_i=1} \geq f(\overrightarrow{x})|_{x_i=0}$*.*

3.2. The importance of monotonicity

Many problems in data analysis can be considered problems in which a collection of data contribute monotonically to some decision. In the conditional independence context, we argue that if two fields within the two databases *do* match, it cannot be the case that we would therefore conclude that the two records *do not* match; for if so this would imply that $\frac{P[x_1=1|M]}{P[x_1=1|\tilde{}M]} < 1$, which would imply that $P\left[x_1 = 1|M\right] < P\left[x_1 = 1|\tilde{}M\right]$. But how can it be that a *field* is *more* likely to match if the *persons* are not the same person? Thus, we argue that the function from the space of possible field matches to the space of record match decisions can in most cases be assumed to be monotonic, and should be modeled as a monotonic boolean function.

3.3. Why the conditional independence model is a linear threshold model

Definition 3.4. *We call a boolean function a "linear threshold" if it can be written in the form* $I\left[\sum_{i=1}^{N} w_i x_i \geq k\right]$ *,where* N *= the number of components;* k *= the chosen threshold;* x_i *= the value of the ith component;* w_i *= the weight associated with the ith component; and I is the "indicator function", taking the value 1 if the statement inside the brackets is TRUE, otherwise taking the value 0 if the statement inside the brackets is FALSE.*

Now, if we temporarily ignore the "don't know" category of the F-S model, then the upper cutoff equals the lower cutoff and the F-S model partitions the data space into two zones, “matched” and “unmatched.” In practice, the mechanism by which this is done is the sum of the logarithms of match weights and nonmatch weights. Under the conditional independence assumption, the F-S model postulates that if the sum of the “match weights” minus the ”nonmatch weights” is greater than the upper cutoff, then the cases match, otherwise they do not. That is, if $\sum_{i=1}^{N} \ln w_i x_i + \ln u_i (1 - x_i) \geq UPPER$ or,$\sum_{i=1}^{N} (\ln w_i - \ln u_i) x_i \geq UPPER - \sum_{i=1}^{N} (\ln u_i)$ where N = number of matching fields; x_i = the result of the comparison on the ith matching field, 1 = matches, 0=nonmatch; w_i = the match weight associated with the ith matching field; u_i = the nonmatch weight associated with the ith matching field; and $UPPER$ = the cutoff for the ”match” threshold in the F-S model, then we declare the two records a match.The CI model fits the definition of a linear threshold: Based on the sum of the results of individual record comparisons, we either add weights (if fields match in the two databases) or subtract nonmatch weights (if they do not) to the sum. Furthermore, we have argued above, and will now assume, that the CI model is monotonic, such that $f(\overrightarrow{x})|_{x_i=1} \geq f(\overrightarrow{x})|_{x_i=0}$. (As noted in [15], this implies that the weights are nonnegative[1].) A linear threshold is a particular kind of map from the space of field comparisons X to the space of record linkage decisions, Y. However, it is not the only kind of map from X to Y. We claim that the most important feature of such a map is that it be *monotonic*, *not* that it be a linear threshold. Further, it is known that the space of linear threshold functions is a subset of the space of monotonic functions—that is, there exist monotonic functions that cannot be written as equal to some linear threshold; but every linear threshold can be written as equal to one of a general class of monotonic functions. Thus, the linear thresholds do not exhaust all possible monotonic functions that we might consider ([15]; [26]); the set of all monotonic functions includes linear thresholds as a proper subset.

4. A nonlinear partial order model based on monotonic boolean functions

We will now outline an alternative, less restrictive and nonparametric model of record linkage. Our central construct will be the monotonic boolean function that maps a space of database fields to a space of linkage decisions. We will begin with a definition of the monotonic function ϕ, whose domain is the vector of field match outcomes, and whose codomain is the decision to accept the record as a

[1] We note that this implication applies only to weights under the CI (no interaction) assumption. Later in this paper, when we add interaction terms, we will find that coefficients associated with interaction terms could be negative and the function would still satisfy monotonicity.

match ($\phi(\overrightarrow{x}) = 1$) or as a nonmatch ($\phi(\overrightarrow{x}) = 0$). We will demonstrate that a monotonic boolean function can be completely characterized by its "minimal path vectors" and "maximal cut vectors", a fact which we shall exploit in the inference section. Note that this section is entirely nonprobabilistic, focusing instead on setting the algebraic foundations for the later statistical inference problem.

Definition 4.1. *Let the state of the decision rule be represented by a binary variable in* $\{0,1\}$. *We let* ϕ *be a discrete function wholly determined by the states of the N members:* $\phi : \{0,1\}^N \rightarrow \{0,1\}$ *will be defined by* $\phi(\overrightarrow{x})$, *with* $\overrightarrow{x} = (x_1, ..., x_N)$. *We call* ϕ *a decision rule. We denote the space* $\{0,1\}^N$ *by* $\mathbb{B}^N$, *and where necessary the structure is fully defined by specification of the pair* $\langle \mathbb{B}^N, \phi \rangle$.

Definition 4.2. *The ith component is irrelevant to* ϕ *if* ϕ *is constant in* x_i; *that is, if* $\phi(\overrightarrow{x})|_{x_i=1} = \phi(\overrightarrow{x})|_{x_i=0}$ *for all combinations of* $\overrightarrow{x}$. *Otherwise the ith component is relevant to the structure.*

Definition 4.3. *A structure* ϕ *is* semi-coherent *if and only if it is monotonic and* $\phi(\overrightarrow{0}) = 0$ *and* $\phi(\overrightarrow{1}) = 1$.

As described in [13], these definitions imply that ϕ defines two sets on the domain of field matching outcomes $\mathbb{B}^N$: The set of vectors $\overrightarrow{x}$ such that $\phi(\overrightarrow{x}) = 1$ (called the on-set, with the set of vectors denoted by Y); and the set of vectors $\overrightarrow{x}$ such that $\phi(\overrightarrow{x}) = 0$ (called the off-set, with the set of vectors denoted by N). These vectors define two concepts, that of a "path vector" (where the record linkage decision is to declare the pair matched) and that of a "cut vector" (where the record linkage decision is to declare the pair non-matched). Additionally, there are minimal path vectors, or min-paths, and maximal cut vectors, or max-cuts, which we now define.

Definition 4.4. *Let* $\overrightarrow{x}$ *indicate the states of the components indexed by* $C = \{0, 1, ..., N\}$. *Define* $C_0(\overrightarrow{x}) = \{i | x_i = 0\}$ *and* $C_1(\overrightarrow{x}) = \{i | x_i = 1\}$. *A path vector is a vector* $\overrightarrow{x}$ *such that* $\phi(\overrightarrow{x}) = 1$. *The corresponding path set is* $C_1(\overrightarrow{x})$. *A cut vector is a vector* $\overrightarrow{x}$ *such that* $\phi(\overrightarrow{x}) = 0$. *The corresponding cut set is* $C_0(\overrightarrow{x})$.

Definition 4.5. *A minimal path vector is a path vector* $\overrightarrow{x}$ *such that* $\overrightarrow{y} < \overrightarrow{x} \Rightarrow \phi(\overrightarrow{y}) = 0$. *The corresponding minimal path set is* $C_1^*(\overrightarrow{x})$. *A maximal cut vector is a cut vector* $\overrightarrow{x}$ *such that* $\overrightarrow{y} > \overrightarrow{x} \Rightarrow \phi(\overrightarrow{y}) = 1$. *The corresponding maximal cut set is* $C_0^*(\overrightarrow{x})$. *If* $\overrightarrow{x}_1 ... \overrightarrow{x}_r$ *is the collection of all min-paths of a semi-coherent structure, then* $C_1^*(\{\overrightarrow{x}_1 ... \overrightarrow{x}_r\}) = \cup_{i=1}^r \{C_1^*(\overrightarrow{x}_i)\}$,*and similarly for* $C_0^*(\overrightarrow{x})$.

For example, consider the structure in which there are 3 components, and we say that any 2 of the 3 components must function for the system to function[2]. Then clearly the set $\{1, 2, 3\}$ indexes a path vector, but the minimal path sets consist of $\{1, 2\}, \{1, 3\}$,and $\{2, 3\}$ (So $C_1^* (\{(1, 1, 0), (1, 0, 1), (0, 1, 1)\}) = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$). The concept of a minimal path vector emerges later in this paper when we are attempting to infer a record linkage rule from data about its field match decisions. For the inference problem, the following proposition is also useful–if a component is relevant to a structure, then it will appear in at least one minimal path set.

Proposition 4.6. *A component x_i is relevant to a structure ϕ if and only if $i \in C_1^* (\overrightarrow{x})$ for some $C_1^* (\overrightarrow{x})$ a minimal path set.*

Proof. See [1] or [15]. ■ Using the concept of a minimal path set, we can write a coherent structure in terms of its minimal paths, as illustrated in the following proposition.

Proposition 4.7. *Let ϕ be a structure function, with a collection of minimal path sets M_j for $j = 1, 2, ..., r$. Let x_{ij} be the ith component in the jth path set. Let $\rho_j (\overrightarrow{x}) = \prod_{i \in M_j} x_{ij}$. Then the coherent structure ϕ can be written as $\phi(\overrightarrow{x}) = \sqcup_{j=1}^{r} \rho_j (\overrightarrow{x})$.*

Proof. See [1] or [15]. ■ Proposition 4.7 sets important groundwork for our later inference problem. The proposition states that if we can construct the minimal path sets of a structure function ϕ, we can write a structure function as a disjunction of all the minimal path sets ($\sqcup_{j=1}^{r} \rho_j (\overrightarrow{x})$) that will be equal to ϕ for all points in $\mathbb{B}^N$. In this sense, then, all of the necessary information about ϕ is "contained" in its minimal path sets. We will exploit this proposition repeatedly in later sections. Furthermore, [13] proves that the upper boundary of the set N, denoted UB(N), is equivalent to the set of "max-cuts" ([1]) of the structure function ϕ, and the lower boundary of the set Y, denoted LB(Y), is equivalent to the set of "min-paths" ([1]) of the structure function ϕ. We define these below, and state, but do not prove the proposition.

Definition 4.8. *Let $\overrightarrow{y}$ and $\overrightarrow{n}$ be vectors in $\mathbb{B}^N$. Define the set $Y = \left\{ \overrightarrow{y} \in \mathbb{B}^N : \phi (\overrightarrow{y}) = 0 \right\}$ as the set of all vectors $\overrightarrow{y}$ such that $\phi (\overrightarrow{y}) = 0$. Similarly, define $N = \left\{ \overrightarrow{n} \in \mathbb{B}^N : \phi (\overrightarrow{n}) = 1 \right\}$ as the set of all vectors $\overrightarrow{n}$ such that $\phi (\overrightarrow{n}) = 1$.*

[2] In the decision context, 2 of the 3 must vote "yes" for the measure to pass and thus the group to vote "yes." In the game theory context, any 2 of the 3 members must form a coalition to win. In the reliability context, 2 of 3 components must function for the system to function. In the diagnosis context, 2 of 3 symptoms must be present for us to diagnose the cause. See [13] for discussion.

Definition 4.9. *(BOUNDARY). Let* N *be a nonempty subset of* $\mathbb{B}^N$ *such that* $\phi(\overrightarrow{n}) = 0$ *for all* $\overrightarrow{n} \in N$*. Define the* upper boundary *of* N*, denoted* $UB(N)$*, as* $UB(N) = \left\{\overrightarrow{n} \in N : \forall \overrightarrow{t} \in N,\ \overrightarrow{n} \geq \overrightarrow{t} \text{ or } \overrightarrow{n} \# \overrightarrow{t}\right\}$*. Similarly, let* Y *be a nonempty subset of* $\mathbb{B}^N$ *such that* $\phi(\overrightarrow{y}) = 1$ *for all* $\overrightarrow{y} \in Y$*. We will define the* lower boundary *of*Y*, denoted*$LB(Y)$*, as* $LB(Y) = \{\overrightarrow{y} \in Y : \forall \overrightarrow{c} \in Y,\ \overrightarrow{y} \leq \overrightarrow{c} \text{ or } \overrightarrow{y} \# \overrightarrow{c}\}$*.*

Remark 1. *Using a standard definition of an ”antichain” in a partial order (e.g., [2], p.95),* $UB(N)$ *and* $LB(Y)$ *are each antichains.*

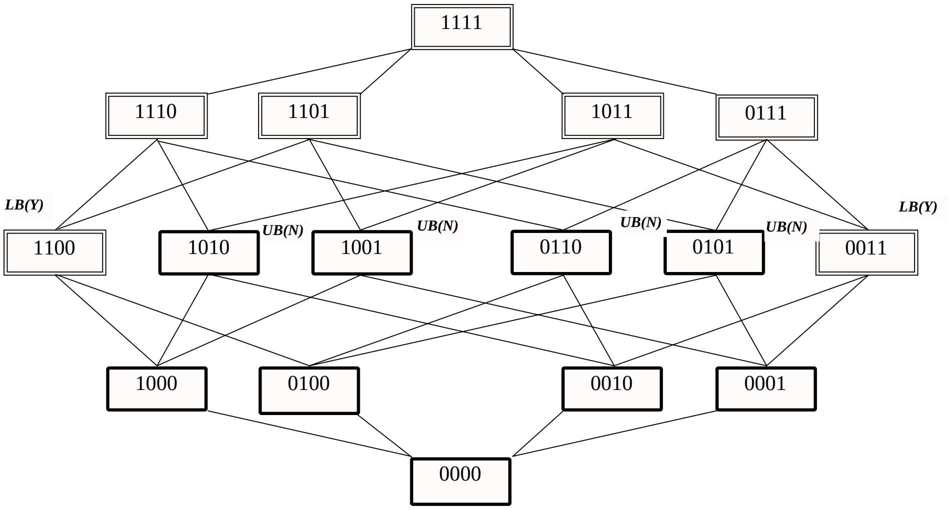
Thus, the upper boundary of a set is a subset of maximal incomparable elements, and the lower boundary of a set is a subset of minimal incomparable elements. Suppose we know the entire structure $\langle \mathbb{B}^N, \phi \rangle$; then $N \cup Y = \mathbb{B}^N$ and the lower boundary of Y and the upper boundary of N have special properties; they are the min-path and max-cut vectors, respectively ([1]).

Proposition 4.10. *Let* $\langle \mathbb{B}^N, \phi \rangle$ *be a semi-coherent structure with minimal path vectors* $\overrightarrow{x}_1, ..., \overrightarrow{x}_r$*, and maximal cut vectors* $\overrightarrow{z}_1, ..., \overrightarrow{z}_s$*, let* $\mathbb{B}^N = N \cup Y$ *and let* $LB(Y)$ *and* $UB(N)$ *be defined as above. Then* $LB(Y) = \{\overrightarrow{x}_1, ..., \overrightarrow{x}_r\}$*, and similarly* $UB(N) = \{\overrightarrow{z}_1, ..., \overrightarrow{z}_s\}$*.*

Proof. See [13]. ■ These three concepts can easily be illustrated by a Hasse diagram for a four component structure, in figure 4.1. The structure for this figure is $\phi(\overrightarrow{x}) = x_1x_2 \sqcup x_3x_4$, with min-path set $\{(1,1,0,0), (0,0,1,1)\}$. We use a square double border box to indicate configurations (that is, $\overrightarrow{x}$'s) in the domain for which $\phi(\overrightarrow{x}) = 1$, and a rounded dark border box to indicate configurations in the domain for which $\phi(\overrightarrow{x}) = 0$.

5. Dealing with stochastic data: A parametric generalized linear model approach

Several authors (see [13], [6], [7], and [28]) focus on the problem of learning a single semi-coherent structure given data about that structure. However, in the record linkage problem, the problem is generalized. Previous authors have imagined a *single* deterministic structure in which we have an active agent deciding which configuration to test next until we have learned the entire structure. We now presume that we are sampling from a population of record linkage decisions (thus our training data set is ”passive” rather than ”active”), and we wish to infer the structure that best represents this population of record linkage decisions. That is, the outcome of the record linkage decision is now a random variable. More formally: Let Y be the record linkage decision, $Y \in \{0, 1\}$, where 1 denotes the



N={(1010),(1001),(0110),(0101),(1000),(0100),(0010),(0001),(0000)} = set of cuts or off-set
UB(N)={(1001),(0110),(1010),(0101)} = set of max-cuts
Y={(1111),(1110),(1101), (1011), (0111),(1100),(0011)} = set of paths or on-set
LB(Y)={(1100),(0011)} = set of min-paths

Figure 4.1: Illustration of upper boundary, lower boundary, min-path set and max-cut set.

decision that the two records match, and 0 denotes the decision that the two records do not match.

Suppose that we construct a $2 \times 2 \times ... \times 2$ cross classification table, where each margin takes on the value 0 or 1 if the corresponding match field is a non-match or a match, respectively. We wish to note that our space $\{0,1\}^N$, and this cross-classification table are equivalent (strictly speaking, there is a set isomorphism between them). We will presume that we have a sampled collection of data in which we know the field match status for each field in the data and the record linkage status for each pair of records in the data. We will presume that all are drawn consistently from a population with a single structure, that is, there exists a correct decision structure on this population. But instead of observing the record linkage structure deterministically, we will assume that we observe the record linkage structure under a stochastic process. We will take our goal to be to infer the best fitting structure from this sampled collection of data coming from that structure.

Our strategy for inferring the best fitting semi-coherent structure from a collection of observations of structures will take the following steps: For each cell in the $2 \times 2 \times ... \times 2$ cross-classification table, we will increment the count in that cell by one if the records are considered a match in our sample of training cases, otherwise we will not increment the count in that cell by one. We will show that, for each semi-coherent structure on $\{0,1\}^N$, there exists a generalized linear model (GLM, see [18]) on the $2 \times 2 \times ... \times 2$ cross-classification table corresponding to the semi-coherent structure. This generalized linear model will incorporate monotonicity constraints. Using data sampled from a population of record linkage decisions, we will find the best-fitting GLM (see [27] for details) for the given cross-classification table, incorporating monotonicity constraints.Finally, we will use the parameter estimates from the best fitting GLM to derive the equivalent semi-coherent structure function, which we will then call our best fitting record linkage rule. In so doing, we will have *parameterized the problem*, transforming it out of its nonparametric form into one where these parameters can be estimated.

Definition 5.1. *Let $\overrightarrow{x}$ be a boolean vector of length N. $\overrightarrow{Y}_{\overrightarrow{x}}$ be a random vector of length 2^N, indexed by $\overrightarrow{x}$. Each element $\overrightarrow{Y}_{\overrightarrow{x}}$ represents a count of the number of decisions in state $\overrightarrow{x}$ in which the records are declared "matched."*

Given a collection of N fields on which we wish to match, there will be 2^N possible configurations of field matches or nonmatches, and thus $\overrightarrow{Y}$ has 2^N elements. For a particular configuration of field match/nonmatch results, we can presume that there exists some probability $p_{\overrightarrow{x}}$ that the record linkage rule will declare that the two records are a match. Again, we index by $\overrightarrow{x}$.

A concrete example will help to illustrate. In figure 4.1, we presumed that the true matching rule was $\phi(\overrightarrow{x}) = x_1 x_2 \sqcup x_3 x_4$, and declare the two records a match if fields 1 and 2 match, OR if fields 3 and 4 match. We, however, are dealing with a probabilistic process, and so, for the configuration $\overrightarrow{x} = (1,1,0,0)$, there is some probability $p_{\overrightarrow{x}}$ that the two records will declared a match and some probability $1 - p_{\overrightarrow{x}}$ that the two records will be declared a nonmatch. Now, if each record linkage decision is independent of the next, then the decisions for the ith configuration will form a sequence of Bernoulli trials, each with probability $p_{\overrightarrow{x}}$ of returning a "match" result. If there are $k_{\overrightarrow{x}}$ cases in the $\overrightarrow{x}$ configuration, and if we count the number of cases returning a match result in the ith configuration and define that random variable as $Y_{\overrightarrow{x}}$, then clearly $Y_{\overrightarrow{x}} \sim BINOMIAL(k_{\overrightarrow{x}}, p_{\overrightarrow{x}})$. This is true for each and every configuration, for all 2^N $\overrightarrow{x}$ configurations. Thus, we make the following definition.

Definition 5.2. *Let $Y_{\overrightarrow{x}} \sim BINOMIAL(k_{\overrightarrow{x}}, p_{\overrightarrow{x}})$, for $k_{\overrightarrow{x}}$ some fixed number of record linkage decisions, and $p_{\overrightarrow{x}} \in [0,1]$.*

By construction, we have proposed that $Y_{\overrightarrow{x}} \sim BINOMIAL(k_{\overrightarrow{x}}, p_{\overrightarrow{x}})$ for each and every of the 2^N $\overrightarrow{x}$ configurations, with the number of trials $k_{\overrightarrow{x}}$ fixed by the sampling design. However, we believe that the probability of the two records being declared a match is a monotonic function of how many, and in what combination, the individual matching fields match.What we need at this point is a link between these two notions.

For generalized linear models, the canonical link function for a dependent binomial random variable and a collection of independent variables is the logit link, $\ln\left(\frac{p_{\overrightarrow{x}}}{1-p_{\overrightarrow{x}}}\right) = \overrightarrow{x}\overrightarrow{\beta}$, which of course implies that $\left(\frac{p_{\overrightarrow{x}}}{1-p_{\overrightarrow{x}}}\right) = e^{\overrightarrow{x}\overrightarrow{\beta}}$, which implies that $p_{\overrightarrow{x}} = \frac{e^{\overrightarrow{x}\overrightarrow{\beta}}}{1+e^{\overrightarrow{x}\overrightarrow{\beta}}}$, which is exactly what we're looking for (Harville and Moore ([10]) developed a similar approach for business linkages). For the remainder of this paper, we shall assume that $p_{\overrightarrow{x}}$ is related to the boolean vector $\overrightarrow{x}$ in such a fashion, and that $\beta_i \geq 0 \; \forall i > 0$.

Definition 5.3. *Let $P(\overrightarrow{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_N x_N + \beta_{12} x_1 x_2 + ... + \beta_{(N-1)N} x_{N-1} x_N + ... + \beta_{123 \cdots N} x_1 \cdots x_N$ (that is, $P(\overrightarrow{x})$ is a linear function of $x_1, ..., x_N$ plus an intercept, plus all $2^N - N - 1$ possible interaction terms), for $\beta_0 \in \mathbb{R}$,and $\beta_1, ..., \beta_{123 \cdots N} \in \mathbb{R}^+$. We will refer to $P(\overrightarrow{x})$ as the* projection *of $\overrightarrow{x}$ into $\mathbb{R}$. In order to avoid confusion, we will use an upper case P for this projection function, and we will use a lower case p for a probability.*

Proposition 5.4. *Suppose $p_{\overrightarrow{x}} = \frac{e^{P(\overrightarrow{x})}}{1+e^{P(\overrightarrow{x})}}$,with $P(\overrightarrow{x})$ as defined above. Then, for any $\overrightarrow{y}$ and $\overrightarrow{z}$ in $\mathbb{B}^N$, if $\overrightarrow{y} \geq \overrightarrow{z}$,then $p_{\overrightarrow{y}} \geq p_{\overrightarrow{z}}$.*

Proof. (Omitted in this version.) ■

Remark 2. *Note: The converse of this proposition is not true;* $p_{\overrightarrow{y}} \geq p_{\overrightarrow{z}}$ *can occur without* $\overrightarrow{y} \geq \overrightarrow{z}$ *if* $\overrightarrow{y} \# \overrightarrow{z}$*.*

This proposition has an immediate corollary showing how to construct structure functions from these $p_{\overrightarrow{x}}$ functions.

Corollary 5.5. *Suppose* $p : \mathbb{B}^N \to [0,1]$ *is defined as* $p\left(\overrightarrow{x}\right) = \frac{e^{P(\overrightarrow{x})}}{1+e^{P(\overrightarrow{x})}}$*. Fix* $r \in [0,1]$ *. Define* $\phi_p\left(\overrightarrow{x}\right) : \mathbb{B}^N \to \{0,1\}$ *by*$\phi_p\left(\overrightarrow{x}\right) = \left\{ \begin{array}{ll} 1 & \mathit{if}\, p_{\overrightarrow{x}} \geq r \\ 0 & \mathit{otherwise} \end{array} \right.$ *.Then* ϕ_p *is monotonic. If, further,* $r \in \left(\min\limits_{\overrightarrow{x}} p\left(\overrightarrow{x}\right), \max\limits_{\overrightarrow{x}} p\left(\overrightarrow{x}\right) \right)$*, then* ϕ_p *is semi-coherent.*

Proof. (Omitted in this version.) ■ Finally, in order to show that we have a relationship between this generalized linear model and the semi-coherent structure ϕ, we must prove that the canonical link we have specified is sufficiently flexible to properly represent *any* semi-coherent structure ϕ.

Theorem 5.6. *Let* ϕ*:*$\{0,1\}^N \to \{0,1\}$ *be any semi-coherent structure function, with collection of min-paths* $M_1, ..., M_r$*. Let* $P\left(\overrightarrow{x}\right)$ *be the projection of* $\overrightarrow{x}$ *into* $\mathbb{R}^+$*. Finally, let* $F\left(\overrightarrow{x}\right) = \frac{e^{P(\overrightarrow{x})}}{1+e^{P(\overrightarrow{x})}}$*,with corresponding* $\phi_F\left(\overrightarrow{x}\right) = \left\{ \begin{array}{ll} 1 & \mathit{if}\, F\left(\overrightarrow{x}\right) \geq \frac{1}{2} \\ 0 & \mathit{otherwise} \end{array} \right.$ *. Then, for any* $\phi(\overrightarrow{x})$*,* $\exists\ 2^N$ *weights* $\beta_0, \beta_1, ...\beta_{x_1 \ldots x_N}$*, with* $\beta_0 \in \mathbb{R}$ *and* $\beta_i \in \mathbb{R}^+$*for* $i = 1, 2, ..., 2^N$*, such that* $\phi(\overrightarrow{x}) = \phi_F\left(\overrightarrow{x}\right), \forall\ \overrightarrow{x} \in \{0,1\}^N$*.*

Proof. (Omitted in this version.) ■ This theorem is the culmination of our model development. We have established that, on a space $\{0,1\}^N$of partially-ordered boolean vectors, on which is defined a semi-coherent structure $\phi\left(\overrightarrow{x}\right)$, we can always find nonnegative real-valued β's by which to construct a function $P\left(\overrightarrow{x}\right)$ that, when transformed by the logistic function $p\left(\overrightarrow{x}\right)$, maps to $(0,1)$,and that this transformation can be further mapped onto a boolean function $\phi_p\left(\overrightarrow{x}\right)$, a decision rule, that is equal to $\phi\left(\overrightarrow{x}\right)$ at all points in the space $\{0,1\}^N$. Our next goal is to show how to estimate such a model with live data.

5.1. Estimation of the model

As specified, our generalized linear model consists of: The structural component $\overrightarrow{x}\overrightarrow{\beta}$; the stochastic specification that $Y_{\overrightarrow{x}} \sim BINOMIAL(k_{\overrightarrow{x}}, p_{\overrightarrow{x}})$, for $k_{\overrightarrow{x}}$

some fixed number of record linkage decisions;the link $p_{\overrightarrow{x}} = \frac{e^{\overrightarrow{x}\overrightarrow{\beta}}}{1+e^{\overrightarrow{x}\overrightarrow{\beta}}}$; and the monotonicity assumption that $\beta_i \in \mathbb{R}^+$ for $i = 1...2^N$. With the exception of the last component, the problem is a standard estimation problem. We deal with the fourth assumption in the next section.

5.2. Incorporating non-negativity constraints: Inequality-constrained maximum likelihood estimation

The nonnegativity constraints on each β_i create a somewhat nonstandard estimation problem. Obviously, applying an unconstrained optimization algorithm on the parameter vector of all possible interactions has the potential to generate negative parameter estimates. Such estimates have been ruled out by the logic of the matching situation, so we must find a way to rule them out in the estimation procedure, as well.

There are several possible options for applying the nonnegativity constraints. We explore two of them. First, computing an unconstrained estimate and applying the Kuhn-Tucker conditions ([17]) to the remaining quadratic programming problem to solve for parameter estimates. This approach has been explored in detail in Judge, et. al. ([12]), and will be explored as a solution method in future work. Second, applying a prior on the β_i's that will incorporate the constraint into the posterior. Also recommended by [12], we might apply a prior to each β_i that is non-negative by definition, and thus will remain non-negative upon incorporation of the data. For example, we might assume a prior on $\beta_i \sim Exp(\theta_i^*)$, for θ_i^* some pre-chosen hyperparameter for each parameter. Obviously, β_i is defined only on the positive real numbers, and hence automatically incorporates our prior information. Finally, one can use the method of projection ([12]:974), which takes a particularly simple form in this context. In the method of projection, any iterative method is used for unconstrained maximization. If, on the nth iteration, an infeasible $\overrightarrow{\beta_n^*}$ is found, it is projected onto a feasible vector $\overrightarrow{\beta_n}$ which is on the boundary of the feasible region. The next maximization step is started from $\overrightarrow{\beta_n}$. In our context, we use the following algorithm: If $\overrightarrow{\beta_n}$ lies within the feasible region ($\beta_i \geq 0$, $\forall i$), stop and return the estimates. If $\overrightarrow{\beta_n}$ is infeasible, find $\beta^* = \min_i \beta_i$, set $\beta^* = 0$.Using iteratively reweighted least squares[3] with $\beta^* = 0$, maximize the unconstrained likelihood function, then go to the first step.The Kuhn-Tucker method is a competitor for future consideration. The Bayesian and Projection methods have the desirable property that, through the projection/maximization

[3] IRLS is recommended by [18] for generalized linear models, and is equivalent to the Gauss method recommended for inequality constrained problems by [12]. IRLS for generalized linear models is asymptotically equivalent to maximum likelihood [18].

process, we learn which combinations of fields are to be ignored (those for which β_i is set to zero or for which the posterior crowds toward zero), and which should be retained (the remainder).

5.3. Finding the implied threshold and interpreting coefficients

After estimating the logistic regression equation, we have estimated parameters $\widehat{\beta_0}, \widehat{\beta_1}, ..., \widehat{\beta_{12...N}}$, all nonnegative. The parameters may be considered voting weights in a nonlinear threshold voting (record linkage) rule, and as estimates of the Fellegi-Sunter likelihood ratios. Obviously, the definitions given above imply that, with some collection of data drawn from the population and obeying the proposed logistic regression relationship, we can estimate β_k for all $k \in \{1, 2, ..., 123...N\}$. But, at this point we ask, what do the individual regression coefficients β_k mean in this context? It is the answer to this question that illustrates how we can use the logistic regression model to estimate Fellegi-Sunter weights.

Theorem 5.7. *For any collection of data drawn from a population satisfying the logistic regression model above, for $k = 2, ..., 123...N$,and all $i \in \{1, 2, ..., K\}$,*

$$e^{\beta_{ik}} \propto \frac{P\left[\textit{the kth configuration is a match, all others nonmatch}|\textit{Records are a match}\right]}{P\left[\textit{the kth configuration is a match, all others nonmatch}|\textit{Records are a nonmatch}\right]},$$

(That is, the regression coefficients are proportional to Fellegi-Sunter weights.)

Proof. (Omitted in this version.) ■ In this context, certain additional information is provided by parameter relationships. Obviously, if we set $\exp(\beta_0) = \frac{P[M]}{P[\tilde{\ }M]}$, we obtain strict equality between β_k and the Fellegi-Sunter weight, so clearly $\exp(\beta_0)$ is an estimate of the prior odds ratio of a record being declared a match under the condition that $x_1 = x_2 = ... = x_{12...N} = 0$. In most populations, we would assume that $\exp(\beta_0)$ is far less than one, indeed it approaches zero. Additionally, we can call $\overrightarrow{F}$ the vector representing some configuration of field match values, $\overrightarrow{F} = (F_1, ..., F_N)$ for $F_i \in \{0, 1\}$, $i = 1, 2..., N$. Since β_k represents the Fellegi-Sunter weight under the configuration $F_k = 1$, $\mathrm{F}_j = 0$ for $j \neq k$, we can use the same methods to derive the following immediate corollary.

Corollary 5.8. *Given any vector of field match values $\overrightarrow{F} = (F_1, ..., F_N)$ for $F_i \in \{0, 1\}$, $i = 1, 2..., N$, $\frac{P[\overrightarrow{F}|M]}{P[\overrightarrow{F}|\tilde{\ }M]} \propto e^{\beta_1 F_1 + ... + \beta_K F_N + ... + \beta_{12...N} F_1 ... F_N}$.*

This corollary implies that, if we have a particular configuration $\overrightarrow{F}$ of match fields and we wish to evaluate their total weight toward a positive record linkage decision, then, consistent with the Fellegi-Sunter model, we merely sum the coefficients associated with fields that match, including all nonzero interaction terms, and exponentiate the sum. Finally, how should the intercept be interpreted? The parallel question is: How to find the threshold? The answer to both questions is given in the following proposition.

Proposition 5.9. *Set the false match and false nonmatch rates equal. After estimating the model* $ln\frac{p}{1-p} = \overrightarrow{x}\widehat{\overrightarrow{\beta}}$, *fix* $p = .5$. *Then* $r = -\widehat{\beta_0}$ *is the optimally predictive threshold for the nonlinear threshold record linkage rule, within the sample data.*

Proof. (Omitted in this version.) ■

6. Illustration of the inequality-constrained maximum likelihood estimation method and the Bayesian method with simulated databases

Following the ideas introduced by Belin ([4]), we designed our simulation as a factorial experiment. We wished to test three different factors affecting the record linkage decision, and determine our ability to reconstruct the true record linkage rule under these different factors in combination. Our first factor is to specify the true decision rule[4]. Our first factor is the true decision rule generating the data: Either a weighted vote rule (we use $3x_1 + x_2 + x_3 + x_4 \geq 3$ as our true decision rule) versus a semi-coherent structure (we use the structure $\phi(\overrightarrow{x}) = x_1x_2 \sqcup x_3x_4$, which cannot be represented as a weighted vote [15]). Our second factor is to identify whether individual matching field results can occur with error. We wish to determine whether the methods proposed here are unduly sensitive to the accuracy of individual field match determinations. Either case 1, where the individual field match results occur with no error; case 2, where the individual field match results occur with a modest amount of error (that is, where the probability that any particular field match is in error is relatively small, we set this probability to .1); or case 3, where the individual field match results occur with a substantial amount of error (that is, where the probability that any

[4] We will note for the record that Winkler, [29], suggests that the belief that there exists a "true" decision rule might be erroneous. He demonstrated that the parameters, cutoffs, and thus, rules vary significantly across geography with seemingly similar files and matching suations of the post-enumeration survey. Developing methods to deal with that complication is a natural direction for future work.

particular field match is in error is relatively large, we set this probability to .2. Our third factor is to specify the properties of the record linkage decision itself. Even given matching fields that are measured without error, we would expect that the record linkage decision in the training sample would be fallible. Our third factor specifies the degree of error for our simulation. Either case 1, where record linkage decisions are made with small error (we set this probability to .05); case 2, where record linkage decisions are made with moderate error (we set this probability to .1); or case 3, where record linkage decisions are made with large error (we set this probability to .2). For the second and third factors, if the field match or record linkage decision is randomly determined to be in error, then the opposite field match or record linkage decision is returned (i.e., "match" becomes "nonmatch" or vice-versa, and "link" becomes "nonlink" or vice-versa). For each level of each factor, we have 20 cases fully crossed with all levels of all other factors. Thus, overall, we have $(20 + 20)\,(20 + 20 + 20)\,(20 + 20 + 20) = 144,000$ individual cases, 8000 in each condition. Our simulation experiment fully crosses these three factors and determines how well the methods described above reconstruct the true decision rule, both under each condition separately and under all conditions together[5].

7. Results of the inequality-constrained method

We first present the results from a selection of conditions. Each condition is labeled, thus condition 1,3,3 represents the condition where the true decision rule is weighted majority, field match decisions are made with large error, and record linkage decisions are made with large error. Similarly, condition 2,3,3 represents the condition where the true decision rule is semi-coherent, field match decisions are made with large error, and record linkage decisions are made with large error. In each case we perform the projection method described earlier, and present the terminating results.

[5] The simulation was also performed with higher levels of error: The results in the second simulation were consistent with those reported here.

7.1. Condition 1,3,3 (Weighted majority, large matching error, large linkage error)

Final glm estimates for condition [1,3,3]

f_3	0.066
	(1.01)
f_1f_2	0.223
	(1.71)
$f_2f_3f_4$	2.249
	(10.10)**
f_1	2.742
	(38.22)**
Constant	-1.394
	(38.07)**
Observations	8000

Absolute value of z-statistics in parentheses
Implied threshold: 1.394
Implied minimal path sets: $\{1\},\{2,3,4\}$

7.2. Condition 2,3,3 (Semi-coherent structure, large matching error, large linkage error)

Final glm estimates for condition [2,3,3]

f_4	0.055
	(0.70)
f_2f_4	0.063
	(0.49)
f_1f_2	2.823
	(23.78)**
f_3f_4	2.882
	(21.96)**
Constant	-1.424
	(41.98)**
Observations	8000

Absolute value of z-statistics in parentheses
Implied threshold: 1.424
Implied minimal path sets: $\{1,2\},\{3,4\}$

8. Results of the Bayesian MCMC method

We now present results of the second method, using Markov-Chain Monte Carlo simulation methods to generate simulated posterior densities. In these simulations, implemented with the software package BUGS ([25]), prior densities for all coefficients except the intercept were chosen to be exponential with a prior hyperparameter of 1. The hyperparameter value of 1 indicates a prior centered on $P\left[x_{ijkl} = 1|M\right] = P\left[x_{ijkl} = 1|\tilde{}M\right]$ indicating a field that provides no information. For the intercept, a normal(0, 166666) prior was used. The large variance simulates very uncertain prior information. After a burn-in of 1000 iterations, 4000 simulated draws from the conditional distributions were performed. Results for two conditions, condition 1,1,3 and condition 2,3,3 are presented below. All other conditions behaved in a fashion similar to these two.

8.1. Condition 1,1,3 (Weighted majority, no matching error, large linkage error)

Final MCMC estimates for condition [1,1,3] with 8000 observations

	Posterior				
Field	mean	std dev	2.5%ile	median	97.5%ile
f_1**	2.718	0.0707	2.58	2.717	2.858
f_2	0.02362	0.02181	6.616E-4	0.01733	0.08125
f_3	0.06474	0.04476	0.003681	0.05694	0.1699
f_4	0.02971	0.02644	8.732E-4	0.02248	0.09847
f_1f_2	0.1707	0.1068	0.01043	0.1545	0.4067
f_1f_3	0.05024	0.04598	0.001407	0.03703	0.1711
f_1f_4	0.04828	0.04457	0.001418	0.03579	0.1699
f_2f_3	0.06624	0.0569	0.001945	0.0516	0.2117
f_2f_4	0.06687	0.05649	0.002479	0.05201	0.213
f_3f_4	0.05014	0.046	0.001233	0.03793	0.1715
$f_1f_2f_3$	0.09006	0.08354	0.00295	0.06441	0.3169
$f_1f_2f_4$	0.1015	0.09135	0.002854	0.07651	0.3384
$f_1f_3f_4$	0.06709	0.06501	0.001556	0.04796	0.2358
$f_2f_3f_4$**	1.982	0.2354	1.53	1.982	2.447
$f_1f_2f_3f_4$	0.1498	0.15	0.004336	0.102	0.5642
Constant	-1.415	0.03524	-1.486	-1.416	-1.345
Observations	8000				

Implied threshold: 1.415

Implied minimal path sets: $\{1\}, \{2, 3, 4\}$

** |Posterior mean / Posterior SD| > 1.96

The posterior kernel densities forβ_0, β_1, β_{234} and β_{14} from the MCMC simulation are displayed below. As expected, the posterior density for β_{14} shifts toward zero, while the intercept β_0, the coefficient on the first field alone (β_1) and the interaction term coefficient (β_{234}) are centered on their posterior estimates.

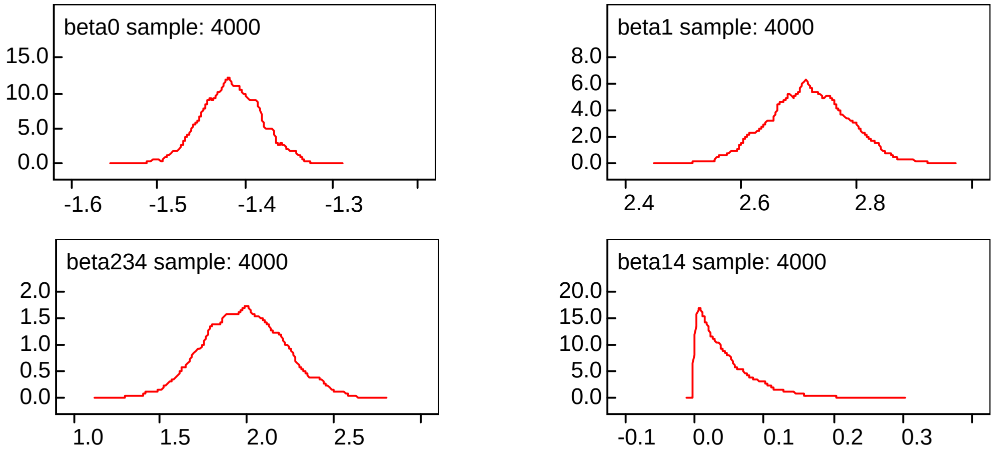


Figure 8.1: Posterior kernel densities for $\beta_0, \beta_1, \beta_{234}, \beta_{14}$ under condition 1,1,3.

8.2. Condition 2,3,3 (Semi-coherent structure, large matching error, large linkage error)

Final MCMC estimates for condition [2,3,3] with 8000 observations

	Posterior				
Field	mean	std dev	2.5%ile	median	97.5%ile
f_1	0.0478	0.0390	0.0018	0.0381	0.1443
f_2	0.0483	0.0385	0.0020	0.0396	0.1434
f_3	0.0459	0.0377	0.0015	0.0362	0.1392
f_4	0.0798	0.0533	0.0044	0.0724	0.2013
f_1f_2**	2.660	0.1262	2.4120	2.6620	2.9080
f_1f_3	0.0610	0.0528	0.0016	0.0469	0.1947
f_1f_4	0.0473	0.0434	0.0014	0.0350	0.1594
f_2f_3	0.0479	0.0433	0.0015	0.0357	0.1611
f_2f_4	0.0881	0.0703	0.0034	0.0715	0.2538
f_3f_4**	2.740	0.1344	2.475	2.740	3.0020
$f_1f_2f_3$	0.1301	0.1141	0.0041	0.1000	0.4291
$f_1f_2f_4$	0.0930	0.0873	0.0026	0.0675	0.3183
$f_1f_3f_4$	0.0835	0.0780	0.0023	0.0608	0.2836
$f_2f_3f_4$	0.1667	0.1442	0.0049	0.1292	0.5295
$f_1f_2f_3f_4$	0.2032	0.2019	0.0054	0.1437	0.7435
Constant	-1.466	0.0355	-1.537	-1.465	-1.398
Observations	8000				

Implied threshold: 1.466

Implied minimal path sets: $\{1,2\}, \{3,4\}$

** |Posterior mean / Posterior SD| > 1.96

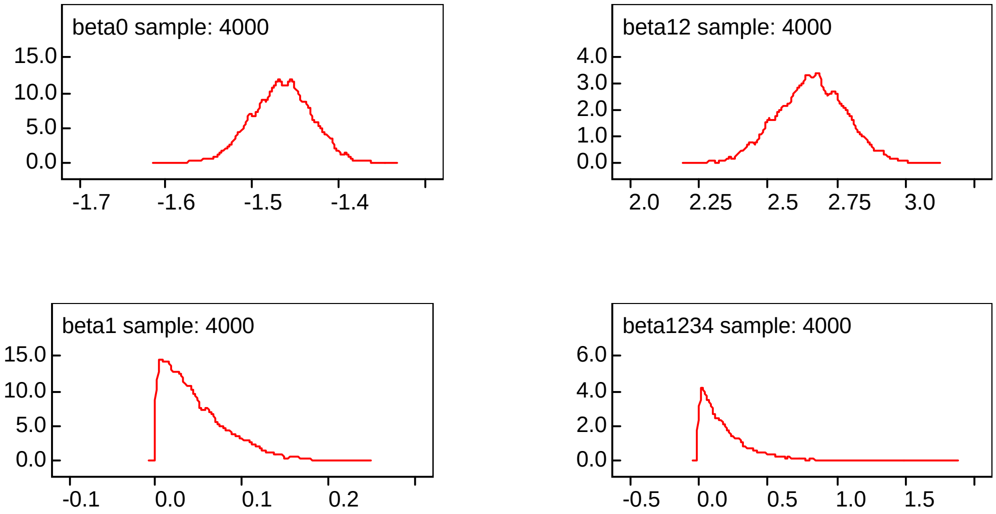


Figure 8.2: Posterior kernel densities for $\beta_0, \beta_1, \beta_{12}, \beta_{1234}$under condition 2,3,3.

The posterior kernel densities forβ_0, β_1,β_{12} and β_{1234} from the MCMC simulation are displayed below. As expected, posterior densities for β_1 and β_{1234} tend to shift toward zero, while the intercept (β_0) and the interaction term coefficient (β_{12}) are centered on their posterior estimates.

9. Conclusions and future research

Based on the ability of the inequality-constrained maximum-likelihood method and the Bayesian method to robustly reconstruct the true decision rule, we conclude that these results are *very* promising. In particular, the ability of the estimation method to find the true rule, merely by focusing on the values of the estimated coefficients of the binomial model, is impressive. These results strongly support further consideration of these methods.

There are several directions future research might take: 1) While the inequality-constrained maximum likelihood and Bayesian approaches are relatively easy to program, the Kuhn-Tucker approach also offers opportunities for development. Both of these should be explored. 2) The Bayesian approach, in particular, would allow us to incorporate information from previous record linkage studies into current record linkage work, thus incrementally improving our ability to infer record linkage rules from new data sets. 3) It has been argued in other contexts that the assumption that we have a training data set ignores some practical limitations in record linkage studies. In particular, record linkage has been used extensively in the context of census and post-enumeration survey (a/k/a A.C.E.) matches, and these matches take place on a very tight time frame. This time frame *might* preclude use of a training data set–although obtaining the training set prior to census/ACE matching might serve to satisfy this objection. However, supposing that a training data set is not available, these methods should be extended into the latent class framework explored by Winkler ([29], [30]), Thibaudeau ([27]), and others. In particular, the inequality-constrained maximum likelihood estimation scheme should be expanded to allow individual record linkage cases to belong to one (matched) or another (unmatched) *latent* class, and the likelihood function should incorporate this model and attempt estimation. 4) An operational test should be performed using this method. To that end, the author and colleagues are developing a test data set of address matches that will be used to compare traditional Fellegi-Sunter methods with the methods described here. The first planned evaluation is simply to compare the methods and their impact on operational matching decisions; later, a test deck will be developed consisting of carefully reviewed "ground truth" cases. From that test deck, evaluations of false positive and false negative error rates will be performed. Evaluating false positive and false negative error rates has been a challenging dilemma for record linkage researchers (see, e.g., [21], [3], [5]), and a general statistical approach has not yet been found to outperform clerical review evaluation methods. 5) The current inequality-constrained maximum likelihood algorithm begins by fitting all possible interactions between individual fields, and eliminates fields one by one using a greedy algorithm. This method suffers two practical disadvantages: First, for N fields, there are 2^N possible coefficients, which grows rapidly in N. Second, a non-greedy algorithm might more rapidly discover the optimal decision rule. Therefore, methods for efficiently exploring the search space are needed. Algorithms used in the data mining literature (e.g., [19]) for bagging, pasting and boosting might be of some help here. In particular, it might be more feasible to begin the search at the intercept-only model, and add field match variables sequentially using some desirable stopping rule, rather than at the fully saturated model and removing terms. Such an approach will be explored in future evaluations of this method.

(References omittted in this version. Available in the full version.)

References

[1] Barlow, R. E., and Proschan, F. (1981). Statistical Theory of Reliability and Life Testing: Probability Models. Silver Spring, MD: To Begin With.

[2] Balakrishnan, V.K. (1995). Shaum's outline of theory and problems of combinatorics. New York, NY: McGraw-Hill.

[3] Belin, T. R. (1991). Using Mixture Models to Calibrate Error Rates in Record-Linkage Procedures, with Applications to Computer Matching for Census Undercount Estimation. Ph.D. Thesis, Harvard University Department of Statistics.

[4] Belin, Thomas R. (1993). Evaluation of sources of variation in record linkage through a factorial experiment. Survey Methodology, 19:13-29.

[5] Belin, Thomas R., and Rubin, Donald B. (1995). A method for calibration of false-match rates in record linkage. Journal of the American Statistical Association, 90:

[6] Boros, E., Hammer, P. L., Ibaraki, T., Kogan, A., Mayoraz, E., and Muchnik, I. (1996). An implementation of logical analysis of data. RUTCOR Research Report RRR 22-96, July, 1996.

[7] Boros, E., Ibaraki, T., and Makino, K. (1996). Boolean analysis of incomplete examples. RUTCOR Research Report RRR 7-96, February, 1996.

[8] Copas, J. and Hilton, F. (1990). Record linkage: Statistical models for matching computer records. Journal of the Royal Statistical Society, Series A, 153:287-320.

[9] Fellegi, Ivan P., and Sunter, A. B. (1969). A theory for record linkage. Journal of the American Statistical Association, 64: 1183-1210.

[10] Harville, D.S., and Moore, R.A. (1999). Determining record linkage parameters using an interative logistic regression approach. Paper presented at the 1999 Joint Statistical Meetings, Baltimore, MD, August 11, 1999.

[11] Jabine, T. (1993). Procedures for restricted access. Journal of Official Statistics, 9:537-590.

[12] Judge, G.G., Griffiths, W.E., Hill, R.C., Lutkepohl, H., and Lee, T-C (1985). The Theory and Practice of Econometrics, 2nd Ed. New York, NY: John Wiley and Sons, Inc.

[13] Judson, D.H. (1999). On the inference of semi-coherent structures from data. Unpublished Master's Thesis, University of Nevada, Reno, Reno, NV.

[14] Judson, D.H. (2000). Estimating Fellegi-Sunter match weights for record linkage using logistic regression. Unpublished paper available from the author.

[15] Judson, D.H., and Gray, L.N. (1994). Coherent structure theory and decision making in social networks: A case study in systems isomorphy. Advances in Group Processes, 11:175-212.

[16] Kmenta, J. (1986). Elements of Econometrics, 2nd Ed. New York, NY: MacMillan.

[17] Kuhn, H., and Tucker, A. (1951). Nonlinear programming, in J. Neyman (Ed.), Second Berkeley Symposium Proceedings. Berkeley, CA: Univ. of California Press.

[18] McCullagh, P., and Nelder, J.A. (1989). Generalized Linear Models, 2nd ed. London: Chapman and Hall.

[19] Ridgeway, Greg (2000). Prediction in the era of massive datasets. Paper presented at the Statistical Modeling for Data Mining conference, Pavia, Italy, October 25-26, 2000.

[20] Robertson, T., Wright, F.T., and Dykstra, R.L. (1988). Order-Restricted Statistical Inference. New York, NY: Wiley.

[21] Rogot, E., Sorlie, P.D., and Johnson, N.J. (1986). Probabilistic methods in matching census samples to the national death index. Journal of Chronic Diseases, 39: 719-734.

[22] Schell, Michael J., and Singh, Bahadur (1997). The reduced monotonic regression method. Journal of the American Statistical Association, 92:128-135.

[23] Scheuren, Fritz, and Winkler, William E. (1993). Regression analysis of data files that are computer matched. Survey Methodology, 19:39-58.

[24] Scheuren, Fritz, and Winkler, William E. (1993). Regression analysis of data files that are computer matched - Part II. Survey Methodology, 23:157-165.

[25] Spiegelhalter, D.J., Thomas, A., and Best, N.G. (1999). WinBUGS Version 1.2 User Manual. MRC Biostatistics Unit.

[26] Taylor, Alan D. (1995). Mathematics and politics: Strategy, voting, power and proof. New York, NY: Springer-Verlag.

[27] Thibaudeau, Yves (1993). The discrimination power of dependency structures in record linkage. Survey Methodology, 19:31-38.

[28] Triantaphyllou, E., Kovalerchuk, B., and Deshpande, A. (1997). Some recent developments of using logical analysis for inferring a boolean function with few clauses. In Barr, R., Helgason, R., and Kennington, J (Eds.), Interfaces in Computer Science and Operations Research: Advances in Metaheuristics, Optimization, and Stochastic Modelling Technologies. Boston, MA: Kluwer Academic Publishers.

[29] Winkler, William E. (1989). Near automatic weight computation in the Fellegi-Sunter model of record linkage. Proceedings of the Fifth Annual Research Conference, U.S. Bureau of the Census, Washington, D.C.

[30] Winkler, William (1995). Matching and record linkage. In: B.G. Cox, et. al., Eds., Business Survey Methods. New York, NY: John Wiley.