Biometrika (1985), **72**, 2, *pp*. 339–51
Printed in Great Britain

A fast procedure for model search in multidimensional contingency tables

By DAVID EDWARDS

Regional Computing Centre at the University of Copenhagen, 2100 Copenhagen, Denmark

and TOMÁŠ HAVRÁNEK

Centre of Biomathematics, Institute of Physiology, Czechoslovak Academy of Sciences, Prague, Czechoslovakia

Summary

A procedure to select the simplest acceptable models for a multidimensional contingency table is proposed. It is based on two rules: first, that if a model is accepted, then all models that include it are considered to be accepted, and secondly, that if a model is rejected, then all its submodels are considered to be rejected. Two versions are described, one for the class of graphical models, and the other for the class of hierarchical log linear models. Application of both versions to a six-dimensional table is illustrated. The procedure can be regarded as an alternative to fitting all possible models, made computationally feasible by application of the two rules. It is a generalization of the procedure proposed by Havránek (1984), but is in many cases considerably faster.

Some key words: Coherence; Contingency table; Distributive lattice; Graphical model; Log linear model; Model selection.

1. Introduction and preliminaries

We consider the analysis of an n-dimensional contingency table N based on a set of classifying factors $\Gamma = A, B, \ldots$. We propose a procedure for selecting parsimonious models for N from two classes of models: $\mathscr{L}_0$, the class of hierarchical log linear models with main effects on N, and $\mathscr{G}_0$, the class of graphical models with main effects.

We describe a model as accepted if it is not rejected by a goodness-of-fit test at some nominal level of significance, otherwise as rejected. We generally use the likelihood ratio test criterion, but this is not essential. Note that for convenience we use the term accepted instead of the more correct expression nonrejected.

The procedure we propose is based on the following rules: first, that if a model is accepted, then all models that include it are considered accepted, and secondly, that if a model is rejected, then all its submodels are considered rejected.

Gabriel (1969) suggested that a test procedure ought not conclude that a model is accepted and another model that includes it is rejected; he termed this principle coherence. Thus the rules just given ensure that this coherence principle is respected.

To clarify the discussion we term a model w-accepted, i.e. weakly accepted, if it includes an accepted model, and w-rejected if it is included in a rejected model.

The procedure searches for a set $\mathscr{A}$ of accepted models and a set $\mathscr{R}$ of rejected models, such that any other model in the class considered either contains a model in $\mathscr{A}$, and so is w-accepted, or is contained in a model in $\mathscr{R}$, and so is w-rejected. Thus all models in the class are accounted for. Some variants of this general procedure are discussed in §3.

The general rationale is as follows. Ideally, all models in the class could be fitted and the simplest models satisfying the criterion be selected. This is however not feasible for higher dimensional tables due to the very large number of possible models. A stepwise procedure can select a simple acceptable model but cannot give information about all models in the class under consideration. In our view it is important to recognize that data may be ambiguous, and that several alternative models may be compatible with the observed table. Thus in contrast to classical hypothesis testing we stress the intention of not missing any hypothesis supported to some extent by the data and which could be important for further study.

The procedure can be regarded as a version of the all possible models approach, in which application of the coherence principle allows the number of models fitted to be very greatly reduced. The set $\mathscr{A}$ selected consists of the simplest models satisfying the given criterion, subject to a proviso described below. These models represent alternative explanations for the table, and give rise to a simple description of which models are compatible with the data; these are precisely the w-accepted models, that is, models that are either in $\mathscr{A}$ or include at least one model in $\mathscr{A}$. This description can be used in different ways according to the purpose of the analysis. Sometimes emphasis may be placed on choosing a single parsimonious model compatible with the data. If there is only one model in $\mathscr{A}$, and it is in accord with subject-matter considerations, or other a priori information, it may be selected. If there are more than one, a choice between them may be made, the choice being based on either a priori information or a more sensitive statistical analysis or both. In other applications it may suffice to draw inference from the set of w-accepted models as a whole, without necessarily adopting any particular model.

The coherence principle is not necessarily satisfied statistically, in the sense that a given goodness-of-fit test may reject a model but accept one of its submodels. In consequence the sets $\mathscr{A}$ and $\mathscr{R}$ found by the procedure are not necessarily unique, and may depend on the choice of steps; see § 3.

We use the term downward search to refer to search going from complex models to simple models, i.e. by removal of edges or model terms. Similarly upward search refers to the opposite direction, i.e. addition of edges or model terms.

Related model selection procedures have been described by Cox & Snell (1974) in the context of linear regression models and by Havránek (1984) for graphical models for contingency tables. Cox & Snell use only the first rule mentioned above and Havránek only the second; see § 3.

2. MODEL REPRESENTATIONS

We can specify a model $m \in \mathscr{L}_0$ by specifying its generating class, $m = [A_1, \ldots, A_p]$, where the A_i are termed generators. For example the model $[AB,BCD]$ for a 4-way corresponds to the log linear expression

$$\log m_{ijkl} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_{ij}^{AB} + \lambda_{jk}^{BC} + \lambda_{kl}^{CD} + \lambda_{jl}^{BD} + \lambda_{jkl}^{BCD}$$

for the expected cell counts m_{ijkl}. The A_i correspond to the maximal terms not set to zero. Here the terms are partially ordered by the inclusion relation applied to the attached set of classifying factors, so that for instance, $\lambda^{AB} \leqslant \lambda^{ABC}$, and maximal and minimal terms are defined with respect to this ordering.

Consider the dual representation of m, written $m = [B_1, \ldots, B_q]^d$, where the B_i

correspond to the minimal terms set to zero, for each value of the indices. Due to the hierarchical nature of the model, this representation is unique. For example, if $\Gamma = \{A, B, C, D\}$, then $[ABC, BCD] = [AD]^d$, $[ABD, ACD, BCD] = [ABC]^d$ and $[BD, AD, CD] = [AB, BC, AC]^d$.

To determine the dual representation of a given model, we may use the following simple algorithm. The B_j are the minimal sets of factors not contained in any A_i. Equivalently, they are the minimal sets that intersect with each $\Gamma \backslash A_i$, $i = 1, \ldots, p$. Thus we can generate all combinations formed by selecting one factor from each $\Gamma \backslash A_i$ for $i = 1, \ldots, p$. The B_i are the minimal sets. For example to determine the dual representation of $[BD, AD, CD]$ we form all possible sets consisting of one factor from each of $\{A, C\}$, $\{B, C\}$ and $\{A, B\}$, giving the sets AB, AC and BC. Thus $[BD, AD, CD] = [AB, AC, BC]^d$.

To determine the usual representation given the dual representation, a similar method may be used. The A_i are the maximal sets that do not include any B_j. Hence $\Gamma \backslash A_i$ are the minimal sets that intersect every B_j. Thus we can generate all sets formed by selecting one factor from B_j, for all j. The A_i are the complements in Γ of the minimal elements in this set. Thus for example, given $[BE, ACE]^d$, we generate the sets BA, BC, BE, EA, EC, E, whose minimal elements are E, BA and BC, so that we obtain $[BE, ACE]^d = [ABCDF, CDEF, ADEF]$.

We use the term dual generators to refer to the generators in the dual representation of a model. Since we only consider main effect models, no dual generators are one element sets. Clearly, $m \in G_0$ if and only if its dual generators are two element sets, that is, correspond to edges.

Models with a single dual generator are also characterized by having generators with $n-1$ elements, when n is the number of factors in the table. Such models have been termed atomic sentences (Havránek, 1982). Let $\wedge$ denote model conjunction, i.e. intersection of the underlying log linear subspaces. Then for any two models written in the dual representation we have that

$$[A_1, \ldots, A_j]^d \wedge [B_1, \ldots, B_k]^d = [A_1, \ldots, A_j, B_1, \ldots, B_k]^d,$$

after removing redundant sets. In particular, for any model

$$[B_1, B_2, \ldots, B_k]^d = [B_1]^d \wedge \ldots \wedge [B_k]^d$$

and this is the canonical representation in atomic sentences suggested by Havránek (1982).

The following formula for model conjunction in terms of the usual representation

$$[A_1, \ldots, A_j] \wedge [B_1, \ldots, B_k] = [A_1 \cap B_1, A_1 \cap B_2, \ldots, A_j \cap B_k]$$

then removing redundant sets, is given, for example, by Havránek (1982). This gives an alternative formulation of the algorithm for obtaining the usual from the dual representation. For example, if $\Gamma = \{A, B, C, D, E\}$, then

$$\begin{aligned}[BE, ACE]^d = [BE]^d \wedge [ACE]^d &= [ACDE, ABCD] \wedge [BCDE, ABDE, ABCD] \\ &= [ABCD, CDE, ADE].\end{aligned}$$

The algorithm for obtaining the dual from the usual representation can similarly be reformulated, as also can the closely related algorithms described in §§3 and 5. We omit the details.

Models are partially ordered by the inclusion relation, written $\leqslant$, that is $m_1 \leqslant m_2$

means that m_1 is a submodel of m_2. If m_1 and m_2 have representations

$$m_1 = [A_{11}, \ldots, A_{1p}] = [B_{11}, \ldots, B_{1q}]^d, \quad m_2 = [A_{21}, \ldots, A_{2r}] = [B_{21}, \ldots, B_{2s}]^d,$$

then clearly $m_1 \leqslant m_2$ if and only if for each A_{1i} there exists an A_{2j} such that $A_{1i} \subseteq A_{2j}$. Using the dual representation, we have $m_1 \leqslant m_2$ if and only if for each B_{2j} there exists a B_{1i} such that $B_{1i} \subseteq B_{2j}$. For example, $[AD, CD]^d \leqslant [AD]^d$ and $[AB, C]^d \leqslant [ABC]^d$. We define maximal and minimal models in respect to this ordering. A minimal model in a set of models S is a model m such that there are no $m' \in S$ satisfying $m' \leqslant m$ and $m' \neq m$. The term maximal is defined similarly.

To represent graphical models we specify which edges are present in the interaction graph, and write accordingly $m = (e_i, e_j, \ldots, e_k)^+$, where $e_i, e_j, \ldots, e_k$ are the edges present in the interaction graph of m. To specify a graphical model in terms of which edges are absent from the interaction graph, we write $m = (e_i, e_j, \ldots, e_k)^-$, where $e_i, e_j, \ldots, e_k$ are the edges absent from the interaction graph. Alternatively we may specify the edges by means of their endpoints, so that, for example, $(AB, BC)^+$ refers to the model with edges AB and BC only.

We conclude this section by relating the model representations described above to lattice theory.

Let for example $\vee$ denote the join operation for models in $\mathscr{L}_0$, that is such that for any two models

$$[A_1, \ldots, A_j] \vee [B_1, \ldots, B_k] = [A_1, \ldots, A_j, B_1, \ldots, B_k]$$

after removing redundant sets. Then $(\mathscr{L}_0, \wedge, \vee)$ forms a distributive lattice, as shown by Havránek (1982).

An element z of a lattice is termed meet-irreducible if $x \wedge y = z$ implies that $x = z$ or $y = z$. The meet-irreducible models in $\mathscr{L}_0$ are precisely the atomic sentences. Any element of a finite distributive lattice has a unique representation as an irredundant meet of meet-irreducible elements (Birkhoff, 1967, Ch. 3), and this corresponds to the dual representation. Similarly the usual representation corresponds to the unique irredundant join of join-irreducible elements. Also $(\mathscr{G}_0, \wedge, \vee)$, where $\wedge$ and $\vee$ are defined in the obvious way by intersection and union of edge sets, forms a Boolean lattice, and hence also a distributive lattice, and the same remarks apply.

3. A model selection procedure for graphical models

We first introduce the concepts of the r-dual and the a-dual of a given set of models. They are used here as the basis of the selection procedure, and may also be generally useful as an aid to structure thinking about the logical relations between models.

Let $S = \{m_1, \ldots, m_p\}, m_i \in \mathscr{G}_0 \ (i = 1, \ldots, p)$ be a set of graphical models and suppose that they have been accepted. Let $I_a(S)$ denote the set of models in $\mathscr{G}_0$ that we can infer are w-accepted, that is,

$$I_a(S) = \{m \in \mathscr{G}_0; m_i \leqslant m \text{ for some } m_i \in S\}.$$

That the models in S are accepted does not imply anything about the remaining models in $\mathscr{G}_0$. These can be written

$$I_a^c(S) = \{m \in \mathscr{G}_0 : m_i \nleqslant m, i = 1, \ldots, p\}.$$

where c denotes complement in $\mathscr{G}_0$.

Let $D_r(S)$ say denote the maximal models of $I_a^c(S)$, that is to say $D_r(S)$ consists of the maximal models in $\mathcal{G}_0$ that do not include any $m_i \in S$. In other words $D_r(S)$ consists of the most complex models which could conceivably be rejected, given the acceptance of the models in S. We call $D_r(S)$ the r-dual of S.

If all the models in the r-dual of S are rejected, then all models in $I_a^c(S)$ are w-rejected, so that every model in $\mathcal{G}_0$ can be designated w-rejected or w-accepted, and the procedure can stop. In this case we call S a complete accepted set.

We define the a-dual of S in a similar way. If all the models in S are rejected, then the set of models that are w-rejected is

$$I_r(S) = \{m \in \mathcal{G}_0; m \leqslant m_i \text{ for some } m_i \in S\},$$

and we define $D_a(S)$, the a-dual of S, to be the minimal models in

$$I_r^c(S) = \{m \in \mathcal{G}_0; m \nleqslant m_i, i = 1, \dots, p\},$$

that is, $D_a(S)$ consists of the minimal models in $\mathcal{G}_0$ that are not included in any model m_i in S. In other words the a-dual consists of the simplest models which could conceivably be accepted, given the rejection of the models in S. If all the models in the a-dual of S are accepted, then the procedure can stop.

Note that $I_a^c(S) = I_r\{D_r(S)\}$: clearly from the definition of $D_r(S)$, we have $I_a^c(S) \subseteq I_r\{D_r(S)\}$. For equality, suppose that for some $m, m \in I_a(S) \cap I_r\{D_r(S)\}$. Then since $m \in I_a(S)$, $m_i \leqslant m$ for some $m_i \in S$, and since $m \in I_r\{D_r(S)\}, m \leqslant m_j^*$ for some $m_j^* \in D_r(S)$. Thus $m_i \leqslant m_j^*$, contrary to definition.

It is clear by construction that when the models in S are incomparable, that is there are no models $m_1, m_2 \in S$ such that $m_1 \leqslant m_2$ and $m_1 \neq m_2$, then

$$D_a\{D_r(S)\} = D_r\{D_a(S)\} = S.$$

This justifies choice of the terms a-dual and r-dual.

To derive $D_r(S)$ for a given set $S = \{m_1, \dots, m_p\}$, write $m_i = (e_{i1}, \dots, e_{in(i)})^+$ $(i = 1, \dots, p)$. Then each model in $D_r(S)$ must omit at least one edge from m_i for each $i = 1, \dots, p$. Thus $D_r(S)$ can be derived as the maximal models in the set

$$(e_{11}, e_{21}, \dots, e_{p1})^-, (e_{12}, e_{21}, \dots, e_{p1})^-, \dots, (e_{1n(1)}, e_{2n(2)}, \dots, e_{pn(p)})^-.$$

Example 1. Consider a 4-way table with classifying factors A, B, C and D, and let $S = \{m_1, m_2\}$, where $m_1 = (AB, BC, AC, CD)^+$ and $m_2 = (AB, BC, AC, BD)^+$. We obtain the models $(AB)^-$, $(AB, BC)^-$, $(AB, AC)^-$, $(AB, BD)^-$, $(BC, AB)^-$, $(BC)^-$, $(BC, AC)^-$, $(BC, BD)^-$, $(AC, AB)^-$, $(AC, BC)^-$, $(AC)^-$, $(AC, BD)^-$, $(CD, AB)^-$, $(CD, BC)^-$, $(CD, AB)^-$, $(CD, BC)^-$, $(CD, AC)^-$ and $(CD, BD)^-$. Thus $D_r(S) = \{m_1^*, m_2^*, m_3^*, m_4^*\}$, where $m_1^* = (AB)^-$, $m_2^* = (BC)^-, m_3^* = (AC)^-$ and $m_4^* = (CD, BD)^-$.

To derive $D_a(S)$ for a given set $S = \{m_1, \dots, m_p\}$, write $m_i = (e_{i1}, \dots, e_{in(i)})^-$ $(i = 1, \dots, p)$. Then each model in $D_a(S)$ must contain at least one of $e_{i1}, \dots, e_{in(i)}$ for each $i = 1, \dots, p$. It follows that $D_a(S)$ can be derived as the minimal models in the set

$$(e_{11}, e_{21}, \dots, e_{p1})^+, (e_{12}, e_{21}, \dots, e_{p1})^+, \dots, (e_{1n(1)}, e_{2n(2)}, \dots, e_{pn(p)})^+.$$

Example 2. Consider again a 4-way table with classifying factors A, B, C and D, and let $S = \{m_1, m_2\}$ where $m_1 = (AB, AC)^- = [BCD, AD]$ and $m_2 = (AB, AD)^- = [BCD, AC]$. Then $D_a(S) = \{m_1^*, m_2^*\}$ where $m_1^* = (AB)^+ = [AB, C, D]$ and $m_2^* = (AC, AD)^+ = [AC, AD, B]$.

Note the close similarity between the algorithms just described for determination of the a-dual and r-dual of a set, and the algorithms concerning dual representation described in §2.

We note that the operation of forming $D_r(S)$ and $D_a(S)$ is associative with respect to the models in S, as illustrated in the following example.

Example 3. Let S be as in Example 1, and let $m_3 = [B, AC, AD] = (AC, AD)^+$. Then $D_r(S \cup \{m_3\})$ is given as the maximal models in the list $(AB, AC)^-$, $(BC, AC)^-$, $(AC)^-$, $(BD, CD, AC)^-$, $(AB, AD)^-$, $(BC, AD)^-$, $(AC, AD)^-$, $(BD, CD, AD)^-$, that is, as $(AC)^-$, $(AB, AD)^-$, $(BC, AD)^-$ and $(BD, CD, AD)^-$.

We can now describe the model selection procedure. This is based on two sets of models, $\mathscr{A}$ and $\mathscr{R}$, containing at any stage the models that have been fitted and were accepted or rejected, respectively.

Step 1. An initial set S_0 of models, which can be chosen freely, is fitted. The models in S_0 are classified into $\mathscr{A}$ or $\mathscr{R}$.

Step 2. If $\mathscr{A}$ is empty, go to Step 2b, and if $\mathscr{R}$ is empty, go to Step 2a. If neither $\mathscr{A}$ nor $\mathscr{R}$ is empty, we can choose between Steps 2a and 2b. Some discussion of this choice is given below.

Step 2a. Fit the models in $D_r(\mathscr{A})\backslash\mathscr{R}$, that is the models in the r-dual of $\mathscr{A}$ that have not already been rejected. If these are all rejected, then stop, otherwise update $\mathscr{A}$ and $\mathscr{R}$ and go to Step 2.

Step 2b. Fit the models in $D_a(\mathscr{R})\backslash\mathscr{A}$, that is the models in the a-dual of $\mathscr{R}$ that have not already been accepted. If these are all accepted, then stop, otherwise update $\mathscr{A}$ and $\mathscr{R}$ and go to Step 2.

We note first that the procedure is bound to terminate after a finite number of steps, since the sets $\mathscr{A}$ and $\mathscr{R}$ increase monotonically. Secondly, we note that it is sufficient to store the minimal models of $\mathscr{A}$ and the maximal models of $\mathscr{R}$, since these determine $I_a(\mathscr{A})$, $I_r(\mathscr{R})$, $D_r(\mathscr{A})$ and $D_a(\mathscr{R})$. Thirdly, the associativity property mentioned above implies that $D_r(\mathscr{A})$ and $D_a(\mathscr{R})$ can be updated directly after each step and need not be calculated afresh from $\mathscr{A}$ and $\mathscr{R}$ each time.

We now discuss the choice of the initial set S_0. Fitting the models formed by removal of one edge from the complete interaction graph was first advocated by Birch (1965), and the associated tests are often referred to as Birch's tests for zero partial association. If the simplest acceptable models have any edges in common, say $e_1, \ldots, e_p$, then it follows that $(e_1)^-, \ldots, (e_p)^-$ must be rejected, so that this choice of S_0 can be interpreted as looking for a possible common core of the simplest acceptable models.

The choice can also be derived in the following manner. The set $\mathscr{A}$ can be considered to be nonempty a priori, since the saturated model is always accepted. If $\mathscr{A}$ contains the saturated model only, when $D_r(\mathscr{A})$ consists of precisely the models formed by removing one edge. Here Step 2a is performed directly, omitting Step 1.

Since the first step may restrict radically the space of models to be considered, it is preferable to make this decision carefully using nonasymptotic tests. Exact tests for zero partial association can be approximated by random sampling; see Table 2 and an unpublished paper of S. Kreiner.

The procedure requires that $I_a(\mathscr{A})$ and $I_r(\mathscr{R})$ have null intersection, which is the case if and only if for no models $m_1 \in \mathscr{A}$ and $m_2 \in \mathscr{R}$ does it hold that $m_1 \leqslant m_2$. If classifying the models in S_0 into $\mathscr{A}$ and $\mathscr{R}$ violates this requirement, then clearly there are several possible simplest acceptable sets, and one must only classify a subset of S_0 consistent with the requirement. There is evidently some advantage in choosing S_0 such that all models in S_0 are incomparable, so that this situation cannnot arise.

The requirement cannot be violated under Steps 2a and 2b. Consider for example fitting a model $m \in D_r(\mathscr{A})\backslash\mathscr{R}$. Then since m is a maximal model of $I_a^c(\mathscr{A})$, we cannot have $m \leqslant m^*$ for $m^* \in \mathscr{R}$, nor can we have $m^{**} \leqslant m$ for $m^{**} \in \mathscr{A}$. Thus, irrespective of whether m is added to $\mathscr{A}$ or $\mathscr{R}$, the requirement cannot be violated.

The natural choice between Step 2a and 2b is to take the step which involves fitting the fewest models, that is compare the number of models in $D_r(\mathscr{A})\backslash\mathscr{R}$ against the number of models in $D_a(\mathscr{R})\backslash\mathscr{A}$. This approach is illustrated in the next section.

We now consider some particular versions of the general procedure. First, we consider the case where S_0 consists of the models formed from the complete interaction graph by removal of an edge, and where Step 2a is chosen each time.

Suppose after the first step $\mathscr{A}$ consists of the models $(e_1)^-, \ldots, (e_p)^-$ and $\mathscr{R}$ consists of $(e_{p+1})^-, \ldots, (e_q)^-$ say. Then $D_r(\mathscr{A})\backslash\mathscr{R}$ consists of $(e_1, e_2)^-$, $(e_1, e_3)^-, \ldots, (e_{p-1}, e_p)^-$ since clearly these models do not include any model in $\mathscr{A}$ and are maximal since the addition of any edge gives a model in $\mathscr{A}$. At the next step $D_r(\mathscr{A})\backslash\mathscr{R}$ consists of models defined by the removal of triplets of edges formed from accepted models of the form $(e_i, e_j)^-$, and so on.

This procedure has been described in detail by Havránek (1984). Note that it only uses the rule that if a model is rejected, all its submodels are considered rejected. It determines the status of all models in $\mathscr{G}_0$ as either accepted or w-rejected. It can be described as a purely downward procedure.

The corresponding purely upward procedure, which only involves use of Step 2b, is as follows. The main effects model is fitted: if it is accepted, the procedure stops, otherwise the models formed by adding an edge to the main effects model are fitted. If models $(e_1)^+, \ldots, (e_p)^+$ say are rejected, then at the next step the models formed from pairs of rejected edges, that is $(e_1, e_2)^+, \ldots, (e_{p-1}, e_p)^+$, and so on. This procedure determines the status of all models in $\mathscr{G}_0$ as either w-accepted or rejected, and it only uses the rule that if a model is accepted, all models that include it are considered to be accepted.

We do not recommend the purely upward procedure, since in most contingency tables we have encountered the main effects model is a long way from the acceptable models, so that this procedure would take too long to reach a solution.

The following modification of the purely upward procedure may be of interest when analysing sparse tables. The same initial set as before, that is the models formed by removing an edge from the complete interaction graph, is fitted. Exact tests may be used, as mentioned previously. Step 2b is then chosen thereafter. Thus this procedure differs from the purely upward procedure in that it starts from the common core of the acceptable models, instead of the main effects model. The models fitted will generally be simpler than with the general procedure, and thus the χ^2 approximations may be more accurate.

4. An example

Table 1 summarizes information collected at the beginnning of a 15 year follow-up study of probable risk factors for coronary thrombosis, comprising data on all men employed in a car factory (Reiniš et al., 1981). There is no clear response structure

Table 1. *Risk factors for coronary heart disease* (*Reiniš et al., 1981*)

			B	no		yes	
			A	no	yes	no	yes
F	*E*	*D*	*C*				
Negative	<3	<140	No	44	40	112	67
			Yes	129	145	12	23
		⩾140	No	35	12	80	33
			Yes	109	67	7	9
	⩾3	<140	No	23	32	70	66
			Yes	50	80	7	13
		⩾140	No	24	25	73	57
			Yes	51	63	7	16
Positive	<3	<140	No	5	7	21	9
			Yes	9	17	1	4
		⩾140	No	4	3	11	8
			Yes	14	17	5	2
	⩾3	<140	No	7	3	14	14
			Yes	9	16	2	3
		⩾140	No	4	0	13	11
			Yes	5	14	4	4

A, smoking; *B*, strenuous mental work; *C*, strenuous physical work; *D*, systolic blood pressure; *E*, ratio of β and α lipoproteins; *F*, family anamnesis of coronary heart disease.

between the six probable risk factors. The models fitted are given in Table 2. The first step consists of removing one edge at a time from the saturated model, i.e. fitting the models $(AB)^-$, $(AC)^-, \ldots, (EF)^-$. At the 5% level, the models $(AC)^-, (AD)^-$, $(AE)^-, (BC)^-$ and $(DE)^-$ are rejected. This can be interpreted as saying that the edges AC, AD, AE, BC and DE must be in any acceptable model. We have

$$\mathscr{R} = \{(AC)^-, (AD)^-, (AE)^-, (BC)^-, (DE)^-\},$$

$$\mathscr{A} = \{(AB)^-, (AF)^-, (BD)^-, (BE)^-, (BF)^-, (CD)^-, (CE)^-, (CF)^-, (DF)^-, (EF)^-\}.$$

The second step consists of fitting the model in $D_a(\mathscr{R})$, that is the single model $(AC, AD, AE, BC, DE)^+$. It is rejected at the 5% level. Thus we obtain

$$\mathscr{R} = \{(AC)^-, (AD)^-, (AE)^-, (BC)^-, (DE)^-, (AC, AD, AE, BC, DE)^+\}.$$

The third step consists of fitting the models in $D_a(\mathscr{R})$. These models contain the edges $AC, AD, AE, BC,$ and DE, plus one edge from

$$\{AB, AF, BD, BE, BF, CD, CE, CF, DE, DF\},$$

that is upward stepping from the model fitted in the second step. Of these only two are not rejected at the 5% level, namely

$$(AC, AD, AE, BC, DE, BE)^+, \quad (AC, AD, AE, BC, DE, CE)^+,$$

corresponding to the addition of the edges BE and CE respectively. Thus we obtain

$$\mathscr{A} = \{(AC, AD, AE, BC, DE, BE)^+, (AC, AD, AE, BC, DE, CE)^+\},$$

since all other accepted models include one of these.

Table 2. *Models fitted in the graphical selection procedure*

Model	d.f.	LR	Asym.	Exact	χ^2	Asym.
$[ACDEF, BCDEF]$	16	22·65	0·1234	0·186	21·21	0·1705
$[ABDEF, BCDEF]$	16	42·80	0·0003	0·001	41·62	0·0005
$[ABCEF, BCDEF]$	16	28·72	0·0259	0·040	27·47	0·0366
$[ABCDF, BCDEF]$	16	40·02	0·0008	0·001	38·76	0·0012
$[ABCDE, BCDEF]$	16	21·31	0·1671	0·255	19·71	0·2336
$[ABDEF, ACDEF]$	16	84·99	0·0000	0·000	631·30	0·0000
$[ABCEF, ACDEF]$	16	12·23	0·7283	0·760	10·99	0·8104
$[ABCDF, ACDEF]$	16	17·23	0·3711	0·452	16·11	0·4451
$[ABCDE, ACDEF]$	16	22·79	0·1195	0·179	23·12	0·1105
$[ABCEF, ABDEF]$	16	14·81	0·5387	0·628	13·34	0·6414
$[ABCDF, ABDEF]$	16	18·63	0·2884	0·391	17·52	0·3529
$[ABCDE, ABDEF]$	16	22·15	0·1383	0·219	20·75	0·1885
$[ABCDF, ABCEF]$	16	31·06	0·0132	0·027	30·22	0·0169
$[ABCDE, ABCEF]$	16	18·35	0·3041	0·394	17·20	0·3730
$[ABCDE, ABCDF]$	16	18·32	0·3057	0·369	17·82	0·3345
$[ADE, AC, BC, F]$	51	83·75	0·0026		81·91	0·0039
$[F, ABC, ADE]$	49	77·77	0·0055		77·78	0·0055
$[BC, AF, ADE, AC]$	50	82·68	0·0025		81·32	0·0034
$[AC, F, BD, BC, ADE]$	50	83·53	0·0021		81·57	0·0032
$[AC, F, BE, BC, ADE]$	50	63·01	0·1023		61·76	0·1229
$[AC, ADE, BF, BC]$	50	79·02	0·0055		77·02	0·0084
$[F, ACD, BC, ADE]$	49	82·99	0·0017		80·86	0·0028
$[F, ACE, BC, ADE]$	49	62·08	0·0994		60·00	0·1349
$[ADE, CF, BC, AC]$	50	83·58	0·0020		81·94	0·0029
$[AC, ADE, BC, DF]$	50	82·63	0·0028		80·16	0·0046
$[AC, ADE, BC, EF]$	50	80·75	0·0040		78·89	0·0060
$[ABCDF, ADEF]$	24	42·56	0·0111		41·11	0·0162

Exact tests were calculated for the first 15 models, with the signifiance level estimated using 1000 random samples; see an unpublished paper by S. Kreiner.

The fourth step consists of examination of $D_r(\mathscr{A})\backslash\mathscr{R}$. This consists of the single model $(BE, CE)^-$, which can be rejected at the 5% level. We can interpret this as meaning that either BE or CE must be in any acceptable model.

We infer that the two models accepted at step three constitute a complete accepted set, and the procedure stops. In the generating set notation, these models are $[AC, ADE, BC, BE, F]$ and $[ACE, ADE, BC, F]$. Their interaction graphs are shown in Fig. 1.

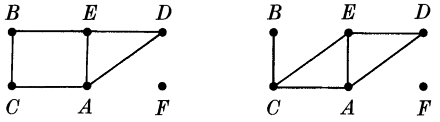

Fig. 1. The interaction graphs of $[AC, ADE, BC, BE, F]$ and $[ACE, ADE, BC, F]$.

We now discuss briefly the interpretation of these models. The most striking feature of both models is the independence of the family anamnesis, F. A dependence on systolic blood pressure, D, and lipoprotein content, E, would be expected since both these factors are partially hereditary in nature. This could be more closely examined in the 3-way table, EDF. Similarly, the absence of direct relation between systolic blood pressure, D, on the one hand, and strenuous physical work, C, and strenuous mental work, B, on the other, merits further investigation.

The models differ in the presence or absence of the edges BE and CE. Lipoprotein content, E, is directly related to strenuous physical work, C, or strenuous mental work, B, or both, but we are unable to decide from the present data which is the case. The ambiguity can be ascribed to a high negative association between strenuous physical and strenuous mental work, in other words, a tendency for work to be either physically or mentally strenuous. A similar phenomenon occurs frequently in regression analysis when covariates are highly correlated.

By fitting a total of 27 models we have obtained information on the w-acceptability or w-rejectability of all $2^{15} = 32768$ possible graphical models under consideration. It is simple to calculate that 768 models are w-accepted and 32000 w-rejected. We infer that the purely downward procedure would require fitting approximately 768 models and the purely upward procedure approximately 32000. The figures are not precise both because different solutions may be obtained and because, for example, the purely downward procedure fits the maximal rejected models in addition to all the accepted models.

5. A model selection procedure for hierarchical models

We now develop the corresponding procedure for the class of hierarchical models with main effects. For a set of models $S = \{m_1, \ldots, m_p\}, m_i \in \mathscr{L}_0$ $(i = 1, \ldots, p)$, we define $I_a(S) = \{m \in \mathscr{L}_0 : m_i \leqslant m$ for some $m_i \in S\}$, so that $I_a^c(S) = \{m \in \mathscr{L}_0 : m_i \nleqslant m, i = 1, \ldots, p\}$. As before we define $D_r(S)$, the r-dual of S, as the set of maximal models in $I_a^c(S)$.

To derive $D_r(S)$ for a given $S = \{m_1, \ldots, m_p\}$, write $m_i = [A_{i1}, A_{i2}, \ldots, A_{in(i)}]$ $(i = 1, \ldots, p)$, omitting generators with one element, corresponding to isolated main effects. Then each model in $D_r(S)$ must set to zero at least one term corresponding to $A_{i1}, \ldots, A_{in(i)}$, for each $i = 1, \ldots, p$. Thus $D_r(S)$ can be derived as the maximal models in the set

$$[A_{11}, A_{21}, \ldots, A_{p1}]^d,\ [A_{12}, A_{21}, \ldots, A_{p1}]^d, \ldots, [A_{1n(1)}, A_{2n(2)}, \ldots, A_{pn(p)}]^d.$$

Example 4. Consider a 4-way table with classifying factors $\Gamma = \{A, B, C, D\}$. Let $m_1 = [ABCD]$ and $S = \{m_1\}$. Then $D_r(S)$ consists of the single model $[ABCD]^d$, that is the model of no 4-factor interaction.

Example 5. Consider again a 4-way table, and suppose $S = \{m_1, m_2\}$, where $m_1 = [BC, ABD, ACD]$ and $m_2 = [BCD, ABD, ABC]$. Then $D_r(S)$ consists of the 5 models $[BC]^d$, $[ABD]^d$, $[ACD, BCD]^d$, $[ACD, ABD]^d$ and $[ACD, ABC]^d$.

Similarly the determination of the a-dual of S can be carried out as follows. Write the models $m_i \in S$ in their dual representation, $m_i = [B_{i1}, \ldots, B_{in(i)}]^d$ for $i = 1, \ldots, p$; then $D_a(S)$, the a-dual of S, consists of the minimal models in

$$[B_{i1}, B_{21}, \ldots, B_{p1}],\ [B_{22}, B_{21}, \ldots, B_{p1}], \ldots, [B_{1n(1)}, B_{2n(2)}, \ldots, B_{pn(p)}].$$

If any of these representations do not contain all factors in Γ, isolated main effects should be appended to them.

The hierarchical selection procedure can now be defined precisely as in the graphical case. The purely upward and purely downward procedures are, however, even less feasible than with graphical models, since there are many more hiearchical models than graphical models. We recommend the same initial set S_0 as in the graphical case, that is models formed by the removal of an edge from the complete interaction graph, since these generally give a fast solution.

6. The example continued

To illustrate use of the hierarchical model selection procedure, we consider again the example of §4. As a starting point we take the sets of accepted and rejected graphical models found at the end of the graphical procedure. We have

$$\mathscr{A} = \{[AC, F, BE, BC, ADE], [F, ACE, BC, ADE]\},$$

$$\mathscr{R} = \{[AC]^d, [AD]^d, [AE]^d, [BC]^d, [DE]^d, [BE, CE]^d\}.$$

with corresponding sets

$$D_r(\mathscr{A}) = \{[AC]^d, [BE, ACE]^d, [BC]^d, [ADE]^d\},$$

$$D_a(\mathscr{R}) = \{[AC, AD, AE, BC, DE, BE, F], [AC, AD, AE, BC, DE, BE, F]\}.$$

Thus both $D_r(\mathscr{A})\backslash\mathscr{R}$ and $D_a(\mathscr{R})\backslash\mathscr{A}$ consist of two models. The results of fitting both sets are given in Table 3.

Table 3. *Models fitted in the hierarchical selection procedure*

Model	d.f.	LR	Asym.	χ^2	Asym.
$[AC, AD, AE, BC, DE, BE, F]$	51	65·85	0·0789	64·63	0·0951
$[AC, AD, AE, BC, DE, CE, F]$	51	64·94	0·0907	62·97	0·1213
$[ABCDF, CDEF, ADEF]$	20	19·02	0·5208	17·91	0·5932
$[BCDEF, ABCEF, ABCDF]$	8	7·48	0·4855	7·03	0·5335

At the 5% level, neither model in $D_a(\mathscr{R})\backslash\mathscr{A}$ can be rejected, and so the procedure stops. The simplest acceptable hierarchical models are $[AC, AD, AE, BC, DE, BE, F]$ and $[AC, AD, AE, BC, DE, CE, F]$, which correspond to the simplest acceptable graphical models found earlier, minus the 3-factor interactions. This is consistent with the test that all three or higher factor interactions are zero (Benedetti & Brown, 1978); see Table 4.

Table 4. *Tests that all $k+1$ and higher interactions are zero*

k	d.f.	LR	Asym. p
0	63	2026·73	0·0000
1	57	843·94	0·0000
2	42	47·35	0·2636
3	22	21·59	0·4823
4	7	9·18	0·2401
5	1	0·41	0·5126

The hierarchical selection procedure may also be used without having recourse to the solution found by the graphical selection procedure. Using the same initial set as given in

Table 2, the procedure finds the same solution after fitting 27 models, the same number as the graphical procedure. The models fitted after the first step are similar, in that edges are replaced by the corresponding 2-factor interactions. In contrast, using the models given in Table 4 as the initial set, the procedure finds the same solution after fitting 38 models in all.

7. DISCUSSION

The procedure appears useful as a preliminary screening method to select simple models compatible with the data, and for this purpose the version for graphical models seems most appropriate (Edwards & Kreiner, 1983). The models selected by the procedure should be regarded as alternative suggestions for the table, and the procedure should be followed by a confirmatory phase, in which the models selected should be examined and compared, both in the light of the subject-matter and also by more sensitive statistical methods. The logical foundations of the conception of data analysis as a process for mechanized hypothesis formation are developed by Hájek & Havránek (1978, Ch. 4).

In our view the virtues of the procedure are that it is efficient, in the sense that only a very small fraction of the models in the class considered need be fitted, and that it is conceptually simple. The computations are not, however, always tractable by hand, as for instance in the example of §4 when a lower significance level, say 0·5%, is used. Finally, we regard it as important that the procedure seeks to identify all the simplest models compatible with the data, and not just one. Thus any ambiguity present in the data is reflected in the results from the procedure.

The weaknesses of the procedure as we see it are primarily weaknesses also inherent in the approach fitting all possible models. When comparing simple nested sequences of models, classical statistical theory rightly emphasises tests for nested hypotheses. These are preferable to individual goodness-of-fit tests, both from a logical viewpoint, in that separate aspects of the models are tested separately, and also because they may have better distributional properties. However, the present approach is not based on simple nested sequences of models, so that tests for nested hypotheses cannot be applied in any straightforward manner.

Similarly fitting all possible models may give incoherent conclusions, and consequently the models selected by the proposed procedure may not be the simplest models satisfying the given criterion. It may be desirable to search for such models, using the selected models as points of departure. We do not regard this as a severe weakness of the procedure, since in our opinion the confirmatory phase should include a sufficiently critical examination of the selected models to reveal any possible further simplification.

Choice of the significance level used is controversial, and no attempt has here been made to study the repeated sampling properties of the procedure. In practice it ought perhaps to be applied using several different levels: in this way the possibility of unnecessary multiple selected models can also be examined.

REFERENCES

BENEDETTI, J. K. & BROWN, M. B. (1978). Strategies for the selection of log linear models. *Biometrics* **34**, 680–6.

BIRCH, M. W. (1965). The detection of partial association, II: the general case. *J.R. Statist. Soc.* B **27**, 111–24.

BIRKHOFF, G. (1967). *Lattice Theory.* New York: American Mathematical Society.

Cox, D. R. & Snell, E. J. (1974). The choice of variables in observational studies. *Appl. Statist.* **23**, 51–9.

Edwards, D. & Kreiner, S. (1983). The analysis of contingency tables by graphical models. *Biometrika* **70**, 553–65.

Gabriel, K. R. (1969). Simultaneous test procedures—Some theory of multiple comparisons. *Ann. Math. Statist.* **40**, 224–50.

Hájek, P. & Havránek, T. (1978). *Mechanising Hypothesis Formation.* Berlin: Springer-Verlag.

Havránek, T. (1982). Some complexity considerations concerning hypotheses in multidimensional contingency tables. In *Transactions of the Ninth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, Ed. J. Kožešník, pp. 281–6. New York: Academic Press; Prague: Academia.

Havránek, T. (1984). A procedure for model search in multidimensional contingency tables. *Biometrics* **40**, 95–100.

Reiniš, Z., Pokorný, J., Basika, V., Tišerová, J., Goričan, K. Horáková, D., Stuchliková, E., Havránek, T., Hrabovský, F. (1981). Prognostický význam rizikového profilu v prevenci ischemické choroby srdce. *Bratis. lek. Listy* **76**, 137–50. (Prognostic significance of the risk profile in the prevention of coronary heart disease).

[*Received February* 1984. *Revised December* 1984]