Computational Statistics & Data Analysis 19 (1995) 155–175
North-Holland


# BIFROST – Block recursive models Induced From Relevant knowledge, Observations, and Statistical Techniques

Søren Højsgaard and Bo Thiesson
*Aalborg University, Aalborg Ø, Denmark*

Received June 1992
Revised July 1993

*Abstract:* The theoretical background for a program for establishing expert systems on the basis of observations and expert knowledge is presented. Block recursive models form the basis of the statistical modelling performed by the program. These models, together with various model selection methods for automatic model selection, are presented. Additionally, the connection between a block recursive model and expert systems based on causal probabilistic networks is treated. A medical example concerning diagnosis of coronary artery disease forms the basis for an evaluation of the expert systems established.

*Keywords:* Causal probabilistic networks; Graphical association models; Machine learning; Model selection; Selection criteria; Selection strategies.

## 1. Introduction

BIFROST is a program for semi-automatic knowledge acquisition and is a continuation of developments made in Greve, Højsgaard, Skjøth and Thiesson (1990). The objective is to obtain preliminary recursive models for use in the HUGIN expert system shell (Andersen, Olesen, Jensen and Jensen 1989) and (Fischer 1990). Based on a database of complete observations on discrete variables and minimal expert guidance the program will search for a model giving a description of the structure of association among the variables. The model obtained can then be transformed into a domain for use in the HUGIN shell.

The model selection is performed in *block recursive models* or *chain graph models*, (see Lauritzen and Wermuth (1984, 1989b)). In block recursive models

*Correspondence to:* B. Thiesson, Institute for Electronic Systems, Aalborg University, Department of Mathematics and Computer Science, Frederik Bajers vej 7, DK-9220, Aalborg Ø, Denmark.


*SSDI* 0167-9473(93)E0054-8

directed and symmetric associations between pairs of variables are combined. A directed association describes an asymmetric dependency in the sense that one variable is a consequence of the other. A symmetric association does not discriminate the associated variables as respectively explanatory and response variable. Hence, the block recursive models cover as special cases the graphical models introduced in Darroch, Lauritzen and Speed (1980), in which all associations are symmetric and the recursive models introduced in Wermuth and Lauritzen (1983), in which all associations are directed.

Block recursive models are appropriate when variables can be arranged in a predefined recursive response structure of blocks (or groups) of variables, where associations between variable pairs in different blocks are directed, whereas associations between variables within the same block are symmetric. For instance, in a medical case variables may be arranged into block(s) of risk factor variables, which precede a block containing disease variables, which again precedes block(s) containing disease manifestations. The practical motivation behind block recursive models is that the relationship between the risk factors should be investigated independently of the diseases and disease manifestations, as these are consequences of the risk factors. Similarly, the relationship between e.g. the diseases and the risk factors should be treated independently of the disease manifestations etc. In the framework of block recursive models these different investigations can then be combined, in a coherent manner, to form one single model.

Hence, model selection in BIFROST demands that the user specifies a block recursive structure by which the types of potential associations between variables are defined. BIFROST will then on the basis of a complete data set determine which associations that actually holds.

The model selection is performed automatically according to specifications given by the user. A model selection method consists of two elements: A selection criterion for selecting among models, and a selection strategy to guide the selection through the set of possible models. As criterion the user can choose between an approach based upon *significance testing*, (see e.g. Whittaker, 1990) or an *information criterion* (as introduced in e.g. Akaike, 1974) – both with various options. Two strategies are available. These are the *direct backward selection strategy* as proposed in Wermuth (1976) and a slightly different strategy denoted the *coherent-direct backward selection strategy*. This strategy, additionally exploits the principle of weak rejection as introduced in the selection strategy presented in Havránek (1984), which in Edwards and Havránek (1985a, 1987a) has been developed further to also include the principle of weak acceptance. The set of possible selectable models can be reduced by specifying prior knowledge of some partial associations which definitely exist and some which definitely do not exist. This specification of prior knowledge influences the quality of the finally selected model.

BIFROST restricts the model selection to *decomposable* block recursive models. This is done for two reasons: Any decomposable block recursive model can be represented by an equivalent recursive model, whereby a selected block

recursive model can be exported to the HUGIN shell. Furthermore, restriction of attention to decomposable models is computationally much more efficient.

A real example concerning diagnosis of coronary artery disease provides a test of some of the selection methods supported by BIFROST. Information criteria have been combined with the coherent-direct backward selection strategy. The expert systems obtained by these selection methods have been evaluated by a reclassification of the data material originally used for the model selection (236 patients) and by a classification of additional 67 patients not part of the original data set. A more detailed evaluation of the selection methods supported by BIFROST will be the subject of a prospective paper. A user's guide to BIFROST can be found in a companion paper (Højsgaard, Skjøth and Thiesson, 1992).

In Section 2 the theoretical basis for block recursive models is described. Section 3 reviews the selection methods supported by BIFROST. In Section 4 the exportation to HUGIN is described. Section 5 describes the medical example. Finally, in Section 6 experiments are described and the results are discussed.

## 2. Block recursive models

In this section the theoretical basis for the statistical modelling in the BIFROST program is presented by outlining the concept and some inference properties of block recursive models. Formally block recursive models generalize graphical models and recursive models. Graphical models play a self-contained role in this context, as the statistical inference relies on a close relationship between a block recursive model and a set of graphical models. For this reason the concept of graphical models and the properties needed are presented as well. Recursive models play a minor technical role due to the ability of BIFROST to export a selected model as a causal probabilistic network into the expert system shell HUGIN.

For a more comprehensive treatment of block recursive models the reader is referred to Lauritzen (1989b), Lauritzen and Wermuth (1989b), Frydenberg and Lauritzen (1989), and Whittaker (1990), whereas graphical models are treated in e.g. Lauritzen (1989a). The reader is referred to e.g. Wermuth and Lauritzen (1983) and Lauritzen, Dawid, Larsen and Leimer (1990) for the concept and a general treatment of recursive models.

### *2.1. Notation and terminology*

*Contingency tables*

Let $\Delta$ be a finite set of discrete random variables. Each variable $\delta \in \Delta$ has a finite set of levels $\mathcal{I}_\delta$. In the contingency table a cell corresponds to an index $i = (i_\delta)_{\delta \in \Delta}$, where $i \in \mathcal{I} = \times_{\delta \in \Delta} \mathcal{I}_\delta$. Let $n$ be the number of observations in the table and let $n(i)$ be the observed counts in cell $i$, $i \in \mathcal{I}$. Then $\Sigma_{i \in \mathcal{I}} n(i) = n$.

Letting $a \subseteq \Delta$ and classifying the observations according to variables in $a$ gives a marginal table with cells $i_a = (i_\delta)_{\delta \in a}$, where $i_a \in \mathcal{I}_a = \times_{\delta \in a} \mathcal{I}_\delta$. The counts in a marginal cell is $n(i_a) = \sum_{j:\, j_a = i_a} n(j)$.

The probability of having an observation in cell $i$ is denoted $p(i)$ where $p(i) \geq 0$ and $\sum_{i \in \mathcal{I}} p(i) = 1$. The probability of an observation in a marginal cell is $p(i_a) = \sum_{j:\, j_a = i_a} p(j)$. It is assumed that the counts in the table are multinomially distributed. Let $A$, $B$, and $S$ denote disjoint sets of random variables. Conditional independence of $A$ and $B$ given $S$ is written $A \perp\!\!\!\perp B \mid S$.

*Graph theory*

Formally a *graph* is a pair $\mathcal{G} = (\Delta, E)$ where $\Delta$ is a finite set of *vertices* and $E$ is a set of *edges* between these. An edge between two vertices $\alpha$ and $\beta$ can be either *directed* (an arrow) from $\alpha$ to $\beta$ or it can be *undirected*, in which case it is referred to as an edge. This is written $\alpha \rightarrow \beta$ respectively $\alpha \sim \beta$. If there is no edge or arrow between $\alpha$ and $\beta$ this is written $\alpha \nsim \beta$. A graph is *undirected* if all edges are undirected, it is *directed* if all edges are directed, and otherwise it is said to be *mixed*. If a directed graph contains no cycles it is said to be *acyclic*.

If there is an arrow from $\alpha$ to $\beta$, then $\alpha$ is a *parent* of $\beta$ and $\beta$ is a *child* of $\alpha$. If there is an undirected edge between $\alpha$ and $\beta$ they are *adjacent* or *neighbours*. The *boundary* of a vertex $\alpha$, $bd(\alpha)$, is the set of parents and neighbours of $\alpha$. A set of vertices $A$ is *ancestral* if $bd(\alpha) \subseteq A$ for all $\alpha \in A$. If two sets $A$ and $B$ are ancestral, then so is $A \cap B$. Hence the *smallest ancestral set* of $A$, $An(A)$, is well defined.

The *ancestors* of $\alpha$, $an(\alpha)$, is the set of vertices from which there is a path (a sequence of distinct edges and arrows) to $\alpha$. If $\alpha \in an(\beta)$ and $\beta \in an(\alpha)$ then $\alpha$ and $\beta$ is said to *connect*, which establishes an equivalence relation, and the corresponding equivalence classes are the *connectivity components* of the graph.

The graph $\mathcal{G}_0 = (\Delta_0, E_0)$ is said to be a *subgraph* of $\mathcal{G} = (\Delta, E)$ if $\Delta_0 \subseteq \Delta$ and $E_0 \subseteq E$. Let $a \subseteq \Delta$ and let $E_a$ denote the set of edges in $E$ between vertices in $a$. Then $\mathcal{G}_a = (a, E_a)$ is the *subgraph induced by a*.

A graph is *complete* if any pair of vertices are connected by an edge or arrow. A subset $a \subseteq \Delta$ is *complete* if the subgraph induced by $a$ is complete, and a complete subset which is maximal w.r.t. inclusion is a *clique*. A subset $S \subset \Delta$ is said to *separate* $A \subset \Delta$ from $B \subset \Delta$ if every path between a vertex in $A$ and a vertex in $B$ contains a vertex from $S$. For an undirected graph $\mathcal{G}$ a triple $(A,B,S)$ of disjoint subsets of $\Delta$ is said to *decompose* $\mathcal{G}$ if $\Delta = A \cup B \cup S$ where $S$ is complete and separates $A$ and $B$.

Let $\Delta$ be partitioned into an ordered set of disjoint subsets, $\Delta(1), \ldots, \Delta(T)$ such that $\Delta = \Delta(1) \cup \cdots \cup \Delta(T)$. A graph is a *chain graph* if edges between vertices in the same sets are undirected whereas edges between vertices in two different sets are directed from the vertex in the set with the lower number to the vertex in the set with the higher number according to the ordering. The connectivity components of the graph are the *chain components* of the graph. Notice that these need not to be identical to the sets $\Delta(1), \ldots, \Delta(T)$ since a set $\Delta(t)$ may contain several chain components.

The *moral graph* of a chain graph $\mathcal{G} = (\Delta, E)$ is the undirected graph $\mathcal{G}^m = (\Delta, E^m)$, where $\alpha \sim \beta$ is an edge in $E^m$ if and only if $\alpha \sim \beta$ or $\alpha \to \beta$ in $\mathcal{G}$ or if there are vertices $\delta_1$ and $\delta_2$ in the same chain component such that $\alpha \to \delta_1$ and $\beta \to \delta_2$. The set $C(t) = \Delta(1) \cup \cdots \cup \Delta(t)$ is the set of *concurrent variables*, and $\mathcal{G}(t)$ is the subgraph induced by these. The graph $\mathcal{G}^*(t)$ is $\mathcal{G}(t)$ with the modifications that the subgraph induced by $C(t-1)$ is completed and all arrows are replaced by edges.

### 2.2. The models

Graphical models and block recursive models are discussed in the following. Notice that we use $\mathcal{G}$ to denote either of undirected or chain graphs. It should, however, be clear from the context which kind of graph we consider.

*Graphical models*

A probability is said to factorise according to an *undirected graph* if there exist non-negative functions $\phi_a$ defined only on complete subsets $a$ such that

$$p(i) = \prod_{a \subseteq \Delta} \phi_a(i_a).$$

If $p(i) > 0$ for all $i \in \mathcal{I}$ and $p$ factorises, $p$ is said to be *Markov*, whereas $p$ is said to be *extended Markov* if $p(i) = \lim_{n \to \infty} p_n(i)$ where $p_n(i)$ are Markov. In the latter case cell probabilities are allowed to be zero.

If $p$ factorises relative to $\mathcal{G}$, $p$ can easily be interpreted in terms of conditional independencies by considering $\mathcal{G}$: If $S$ separates $A$ and $B$ in the graph, then $A \perp\!\!\!\perp B \mid S$ and, as a special case, if $\alpha \not\sim \beta$, then $\alpha \perp\!\!\!\perp \beta \mid \Delta \setminus \{\alpha, \beta\}$. These properties are known as the *global Markov property* respectively the *pairwise Markov property*, and provide an easy tool for interpreting a model.

A *model* $\mathcal{M}$ is a family of probabilities satisfying the same set of conditional independencies. If these are depicted in a graph $\mathcal{G}$ we write $\mathcal{M}(\mathcal{G})$. Let $\bar{\mathcal{M}}(\mathcal{G})$ denote the set of extended Markov probabilities relative to $\mathcal{G}$. Formally the set of probability distributions $\bar{\mathcal{M}}(\mathcal{G})$ is a *graphical model*.

An important subclass of the graphical models are the *decomposable models*. These have the property that the probability distribution can be factorised successively in accordance with a decomposition of the graph. It can be shown (see Lauritzen, Speed and Vijayan, 1984) that a graphical model is decomposable if its graph contains no cycles of length $\geq 4$ without a chord.

*Block recursive models*

To define a block recursive model consider a probability $p$ and a chain graph $\mathcal{G}$. If for any two non-adjacent vertices $\alpha$ and $\beta$ it holds for $p$ that

$$\alpha \perp\!\!\!\perp \beta \mid C(t^*) \setminus \{\alpha, \beta\},$$

where $t^*$ is the smallest $t$ such that $\{\alpha, \beta\} \in C(t)$, $p$ is said to satisfy the *block recursive Markov property*. If $p$ is strictly positive it can be shown (Frydenberg,

1990a) that this property does not depend on a particular partitioning of the variables into the blocks, but only relies on the graph itself. (Recall, that $\Delta(1), \ldots, \Delta(T)$ are *not* necessarily the chain components of the graph.) If $p$ is strictly positive the block recursive Markov property is equivalent to the *global chain Markov property*: If $S$ separates $A$ and $B$ in $(\mathcal{G}_{An(A \cup B \cup S)})^m$, the moral graph of the smallest ancestral set containing $A$, $B$ and $S$, then $A \perp\!\!\!\perp B \mid S$. This Markov property makes the interpretation of the model straight forward.

Formally a *block recursive model* $\bar{\mathcal{M}}(\mathcal{G})$ is a family of *extended Markov probabilities* relative to $\mathcal{G}$, i.e. either strictly positive distributions satisfying the block recursive Markov property or limits of such distributions. By this cell probabilities are allowed to be zero and the global chain Markov property will still remain valid.

There is a close relationship between block recursive models and graphical models. Consider a factorization of a distribution relative to the partitioning of a chain graph $\mathcal{G}$ into blocks of variables $\Delta(1), \ldots, \Delta(T)$. Then

$$p(i) = \prod_{t=1}^{T} p\left(i_{\Delta(t)} \mid i_{C(t-1)}\right) = \prod_{t=1}^{T} \frac{p\left(i_{\Delta(t) \cup C(t-1)}\right)}{p\left(i_{C(t-1)}\right)}. \tag{1}$$

Recall that $\mathcal{G}^*(t)$ is the subgraph induced by $C(t)$ with the modifications that the subgraph induced by $C(t-1)$ is made complete and that all arrows are replaced with edges. It can be shown (Lauritzen, 1992) that $p$ satisfies the extended Markov property relative to $\mathcal{G}$ if and only if any numerator in the rightmost expression in (1) is extended Markov with respect to $\mathcal{G}^*(t)$. Thus the independence structure among the variables reflected in the chain graph $\mathcal{G}$ are also reflected in the undirected graphs $\mathcal{G}^*(t)$, $t = 1, \ldots, T$.

Let $\bar{\mathcal{M}}(\mathcal{G}^*(t))^C$ denote the graphical model relative to $\mathcal{G}^*(t)$ conditional on the variables $C(t-1)$. Each term $p(i_{\Delta(t)} \mid i_{C(t-1)})$ of the factorization (1) is the conditional distribution of $\Delta(t)$ given $C(t-1)$ relative to $\mathcal{G}(t)$, and describes the structure of association between any pair of variables $\{\alpha, \beta\}$ where $\alpha \in \Delta(t)$ and $\beta \in C(t)$. Hence, $p(i_{\Delta(T)} \mid i_{C(t-1)}) \in \bar{\mathcal{M}}(\mathcal{G}^*(t))^C$. Further, the restrictions on these terms are independent since they relate to different sets of variables, and the equivalence between a block recursive model and graphical models on $\mathcal{G}^*(1), \ldots, \mathcal{G}^*(T)$ then gives the factorization of the block recursive model into terms of graphical models:

$$\bar{\mathcal{M}}(\mathcal{G}) = \bar{\mathcal{M}}(\mathcal{G}^*(1))^C \times \cdots \times \bar{\mathcal{M}}(\mathcal{G}^*(T))^C. \tag{2}$$

Finally, in accordance with this factorization a block recursive model is said to be *decomposable* if all $\mathcal{G}^*(t)$ are decomposable.

### 2.3. Estimation

For a graphical model relative to an undirected graph $\mathcal{G}$ with the set of cliques $\mathcal{C}$, the maximum likelihood estimate is determined uniquely by the set of equations $\hat{p}(i_c) = n(i_c)/n$, $c \in \mathcal{C}$, $i_c \in \mathcal{I}_c$. In general these equations must be

solved iteratively, e.g. by successively fitting the marginals by the IPS algorithm (see e.g. Lauritzen, 1989a). If, however, the model is decomposable, the maximum likelihood estimate can be found explicitly, i.e. without iteration. Thereby the required computational effort is reduced considerably.

For a chain graph the likelihood function factorises according to the partitioning of the variables. Under the assumption of multinomial sampling the likelihood function becomes

$$\begin{aligned} L(p) &\propto \prod_{i\in\mathcal{I}} p(i)^{n(i)} = \prod_{i\in\mathcal{I}} \prod_{t=1}^{T} p\left(i_{\Delta(t)} \mid i_{C(t-1)}\right)^{n(i)} \\ &= \prod_{t=1}^{T} \prod_{i_{C(t)}\in\mathcal{I}_{C(t)}} p\left(i_{\Delta(t)} \mid i_{C(t-1)}\right)^{n(i_{C(t)})} \\ &= \prod_{t=1}^{T} L_t(p_t). \end{aligned} \tag{3}$$

Since $p(i_{\Delta(t)} \mid i_{C(t-1)})$, $t = 1, \cdots T$ are mutually independent, the likelihood functions $L_t(p_t)$ can be maximized independently. In practice, the maximum likelihood estimate for $p_t$ can be found as

$$\hat{p}\left(i_{\Delta(t)} \mid i_{C(t-1)}\right) = \frac{\hat{p}^*\left(i_{\Delta(t)\cup C(t-1)}\right)}{\hat{p}^*\left(i_{C(t-1)}\right)} = \frac{\hat{p}^*\left(i_{C(t)}\right)}{n\left(i_{C(t-1)}\right)/n}, \tag{4}$$

where the term $\hat{p}^*$ is the maximum likelihood estimate for $p \in \bar{\mathcal{M}}(\mathcal{G}^*(t))$. This follows from proposition 5 in Frydenberg and Lauritzen (1989) by observing that $(\emptyset, \Delta(t), C(t-1))$ decomposes $\mathcal{G}^*(t)$. See also Frydenberg (1990b). The simple denominator in the right most expression in (4) is obtained by observing that the subgraph $\mathcal{G}^*(t)$ induced by $C(t-1)$ is complete.

### 2.4. Inference

The deviance between two models and the dimension of a model are central quantities for the statistical inference in block recursive models in this paper. It turns out that inference in block recursive models can be made equally well in appropriate undirected graphical models, and that the inference in this set of undirected graphical models can be performed independently.

Let $\mathcal{G}$ and $\mathcal{G}_0$ be chain graphs where $\mathcal{G}_0 \subseteq \mathcal{G}$, and let $\hat{p}$ and $\hat{p}_0$ be the maximum likelihood estimates for the models $\bar{\mathcal{M}}(\mathcal{G})$ and $\bar{\mathcal{M}}(\mathcal{G}_0)$. According to (3) the factorization of the probability relative to the block recursive structure of the graphs gives the likelihood ratio

$$Q = \prod_{t=1}^{T} \left( \prod_{i_{C(t)}\in\mathcal{I}_{C(t)}} \frac{\hat{p}_0\left(i_{\Delta(t)} \mid i_{C(t-1)}\right)}{\hat{p}\left(i_{\Delta(t)} \mid i_{C(t-1)}\right)} \right)^{n(i_{C(t)})} = \prod_{t=1}^{T} \left( \prod_{i_{C(t)}\in\mathcal{I}_{C(t)}} \frac{\hat{p}_0^*\left(i_{C(t)}\right)}{\hat{p}^*\left(i_{C(t)}\right)} \right)^{n(i_{C(t)})},$$

where the right most term is obtained from the relationship to estimation in

graphical models stated in (4). Thereby the deviance (defined as $-2 \log Q$) becomes additive over the blocks

$$D = \sum_{t=1}^{T} 2 \sum_{i_{C(t)} \in \mathcal{I}_{C(t)}} n(i_{C(t)}) \log \frac{\hat{p}^*(i_{C(t)})}{\hat{p}_0^*(i_{C(t)})} = \sum_{t=1}^{T} D_t.$$

From (2) it follows directly that the dimension of a block recursive model, i.e. the number of parameters, is also additive over the blocks,

$$\dim(\bar{\mathcal{M}}(\mathcal{G})) = \sum_{t=1}^{T} \dim(\bar{\mathcal{M}}(\mathcal{G}^*(t))^C) = \sum_{t=1}^{T} \dim_t,$$

implying that the the degrees of freedom between two models become additive over the blocks as well.

The dimension of $\bar{\mathcal{M}}(\mathcal{G}^*(t))^C$ can be found by exploiting again that $(\emptyset, \Delta(t), C(t-1))$ forms a decomposition of $\mathcal{G}^*(t)$. Hence, by proposition 4 in Frydenberg and Lauritzen (1989)

$$\dim(\bar{\mathcal{M}}(\mathcal{G}(t))^C) = \dim(\bar{\mathcal{M}}(\mathcal{G}^*(t))) - \dim(\bar{\mathcal{M}}(\mathcal{G}^*_{C(t-1)}(t))). \tag{5}$$

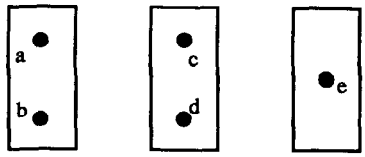
The dimensions of graphical models on the right hand side of (5) are easily obtained by means of e.g. the recursion formulas as presented in Lauritzen (1989a, p. 32).

The factorization of a block recursive model into sets of mutually independent graphical models (2) and the additivity of the deviance and degrees of freedom according to these models imply that model selection within block recursive models can be performed as independent model selections within the sets of associated graphical models. Notice, by the definition of $\bar{\mathcal{M}}(\mathcal{G}^*(t))^C$ that the lattice of graphical models associated with block $\Delta(t)$ only differs on independencies implied by exclusion of edges involving at least one node in $\Delta(t)$. The trick to base the model selection on the associated graphical models is illustrated in Figure 1.

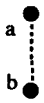
## 3. Model selection

The purpose of a model selection process is to select the model containing the nearest representation of the actual problem which it is possible to construct by means of a probability distribution.

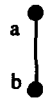
In order to select this "ideal" model from a family of models, ideally the suitability of all possible models should be checked according to a chosen criterion and the most satisfying model then selected. Naturally, this is not practically possible within a family of graphical models, except for trivial examples. The number of graphical models with $n$ variables is given as $\sum_{i=0}^{n} \binom{n}{i} 2^{\binom{i}{2}}$. Accounting only for decomposable graphical models does reduce this number perceptibly, but still the number of models is very large. (The



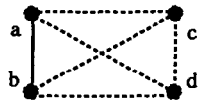

1) Block recursive structure



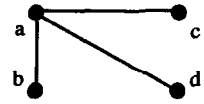

2a) Selectable models for block 1



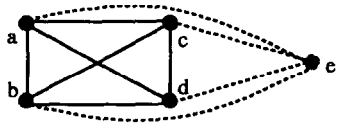

2b) Selected model for block 1



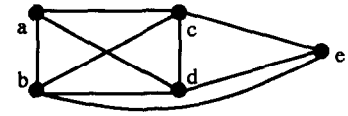

3a) Selectable models for block 2



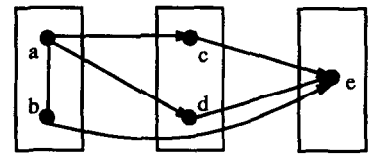

3b) Selected model for block 2




4a) Selectable models for block 3




4b) Selected model for block 3




5) Selected block recursive model

Fig. 1. Model selection in block recursive models based on associated graphical models. The block recursive structure contains 3 blocks, as illustrated in 1). In 2a), 3a), and 4a) the lattices of graphical models associated with each block are illustrated. An edge is dashed if the edge may be removed to construct a representation of a simpler graphical model. 2b), 3b), and 4b) represent a final selected model for each block. In 5) the selected graphical models associated with each block are combined into the final selected block recursive model.

number of decomposable graphical models cannot be determined by a closed expression (Lauritzen, 1989a)). Hence, a strategy reducing the search space considerably must be imposed.

Considering this, a model selection method consists of two elements: A selection criterion which serves to judge the suitability of a model, and a

selection strategy which serves to guide the selection process through a set of possible models.

### 3.1. Selection criteria

In BIFROST a traditional *significance test criterion* can be applied. The concept of significance testing is based on applying a test statistic to judge whether the set of observations are exceptional under a given hypothesis. In that case the hypothesis is rejected due to inconsistency with the observations. For significance testing the smallest model which cannot be rejected constitutes the "ideal" model.

In BIFROST the deviance is used as test statistic and the evaluation can be performed by determination of $p$-values relative to a $\chi^2$-distribution or by simulated exact deviance test using the algorithm of Patefield (Patefield, 1981).

The $\chi^2$-test approach is based on the fact that the deviance between two comparable models is asymptotically $\chi^2$-distributed with degrees of freedom equal to the difference in their dimensions. This is a large sample approximation, which is often not satisfied due to sparseness caused by high dimensional models. In essence, a simulated exact test is based on simulating a number of tables with the same sufficient marginals as those observed. Then the deviance is evaluated for the observed as well as the simulated tables and the proportion of deviances larger than the observed deviance is then used as an estimate for the $p$-value. The simulated exact test is only applicable within decomposable models.

As an alternative to significance test BIFROST allows application of *information criteria*. The general notion of information criteria is developed in accordance with the structural form of the Akaike information criterion (Akaike, 1974). This latter criterion is given as an estimate of the expected Kullback-Leibler information divergence between the fictitious true distribution and the estimated distribution of an element from the set of possible models. Since the true distribution is a conceptual notion an approximation is performed by replacing the true distribution with the sample distribution. The Akaike information criterion $AIC$ is in Akaike (1974) formulated as

$$AIC = D + 2\ dim,$$

where $D$ is the deviance between the saturated model and the considered model, and $dim$ is the dimension of the considered model.

As an alternative to $AIC$ we suggest a heuristic information criterion ($IC$) given as

$$IC = D + \kappa\ dim,$$

where $\kappa$ is a penalty parameter for the complexity of the model to be specified by the investigator(s).

The information criteria can be given a reasonable interpretation. The deviance $D$ is a measure of the fit of the estimated distribution of the model to the observations, and the dimension dim of the model is a measure of the

complexity of the considered model. Consequently, the use of an information criterion can be seen as a trade-off between fit and complexity guided by the penalty parameter $\kappa$. If the estimated distribution of the considered model is very close to the sample distribution, then the fit is good reflected by a relatively small deviance. This model, however, might be quite complex, for which reason the number of parameters is large, and vice versa. The model with the smallest *IC* value constitutes the "ideal" model using the information criterion.

### 3.2. Selection strategies

The strategy approaches chosen in BIFROST is a *stepwise backward elimination strategy* and a variant of this, additionally utilizing the principle of coherence, to be described below. The stepwise backward elimination strategy has been proposed by Goodman (1971) and adapted to (decomposable) graphical models by Wermuth (1976).

Consider a model $\bar{\mathcal{M}}(\mathcal{G}_i)$, and let $S[\bar{\mathcal{M}}(\mathcal{G}_i)]$ denote the set of maximal models $\bar{\mathcal{M}}(\mathcal{G}_i^j)$, $j = 1, \ldots, k$ which by the partial ordering of models are simpler than $\bar{\mathcal{M}}(\mathcal{G}_i)$. Thus $S[\bar{\mathcal{M}}(\mathcal{G}_i)]$ is the set of models for which the graphical representations $\mathcal{G}_i^j$ can be obtained from $\mathcal{G}_i$ by removal of exactly one edge.

The general idea of the strategy is as follows. Starting from the saturated model, the final model is selected by a stepwise local "ideal" determination of which conditionally independencies between pairs of variables can be assumed given the evidence of the data.

At each step the set $S[\bar{\mathcal{M}}(\mathcal{G}_i)]$ is determined. Assuming $S[\bar{\mathcal{M}}(\mathcal{G}_i)] \neq \emptyset$ a model $\bar{\mathcal{M}}(\mathcal{G}_i^{j*}) \in S[\bar{\mathcal{M}}(\mathcal{G}_i)]$ with the optimal value according to a chosen criterion is decided. For significance test the optimal value equals the maximum $p$-value. For *IC* the optimal value equals the minimal value. Given that the $p$-value of $\bar{\mathcal{M}}(\mathcal{G}_i^{j*})$ surpasses the critical value when significance test criteria are used or given that the *IC* value of $\bar{\mathcal{M}}(\mathcal{G}_i)$ subtracted the *IC* value of $\bar{\mathcal{M}}(\mathcal{G}_i^{j*})$ is non-negative when *IC* criteria are used, then $\bar{\mathcal{M}}(\mathcal{G}_i^{j*})$ is selected as initializing model for the next step. That is, the variable pair with the smallest partial association is decided to be conditionally independent. If $\bar{\mathcal{M}}(\mathcal{G}_i^{j*})$ does not obey this condition or if $S[\bar{\mathcal{M}}(\mathcal{G}_i)] = \emptyset$, then $\bar{\mathcal{M}}(\mathcal{G}_i)$ becomes the final selected model.

Naturally, any graphical model may be derived by a stepwise elimination of edges from the saturated model. A formal proof that this property holds true within a family of decomposable models has been given by Edwards (1984). The model selection by stepwise elimination can therefore be restricted to the decomposable models without leaving any models unattainable.

An alternative to the stepwise approach is a reduced version of the all possible models approach. Two different types of this approach have been proposed by respectively Edwards and Havránek (1985b, 1987b) and Madigan and Raftery (1991). To reduce the number of models checked, the proposed strategies rely on the following principles:

- *Weak rejection principle:* If a model is rejected, then all of its submodels can be considered rejected.

- *Weak acceptance principle:* If a model is accepted, then all models that include it can be considered accepted.

The principles were introduced by Gabriel (1969), where the principles regarded as properties of a selection criterion were attached the term *coherence*.

It is easy to prove that if a selection criterion is based exclusively on the deviance it is coherent. However, for significance testing where the deviance is compared to a critical value depending on degrees of freedom, the criterion is non-coherent. Also, criteria based on *IC* values are non-coherent as the two involved effects, the deviance and the number of parameters, act in opposition to each other. Hence, when coherence is used to restrict the set of checked models when significance test or *IC* are used as criterion, coherence becomes a principle rather than a property. The principle then relies on the presuposition that only occasionally we may find that the more general model of two comparable models is rejected, whereas the simpler is accepted by the criterion.

Now, consider model selection within large families of models. If a strategy is based exclusively on the principle of coherence, the starting model [1] becomes of crucial importance. If early rejections of large models and/or early acceptance of small models are not experienced the strategies become extremely "expensive" due to absence of extensive reductions in the remaining set of models [2].

As mentioned, we have, however, utilized the principle of coherence in BIFROST. Additionally to the stepwise backward elimination strategy, called the *direct backward strategy*, we have implemented a strategy additionally utilizing the principle of weak rejection. This strategy is called the *coherent-direct backward strategy*. The coherent-direct strategy differs from the direct strategy in the way that at each step only models not already rejected by the weak rejection principle are competing for the selection as initializing model for the next step. Hence, this hybrid strategy ensures relatively fast selection of a final model due to the stepwise local "ideal" selection, and depending on earlier rejections, efficiency is gained at each step by only considering models which are not already weakly rejected.

Unfortunately, the principle of stepwise local "ideal" selection attaches some risks of misselection. It is not certain that stepwise local "ideal" selections will end up in a global "ideal" selection due to detour(s) forced by earlier local selections. Also, as described earlier, the coherent-direct strategy applied with a non-coherent criterion enforces the risk that the global "ideal" model will be cut off by weak rejection.

[1] Strategies based exclusively on the principle of coherence do not necessarily start from a single model. They may be started from a set of incomparable models. Similarly, these strategies may reach a set of incomparable final selected models instead of a single model. For details on this, see the above mentioned references.

[2] On the example described in Section 5 we have experimented with a backward strategy based on the weak rejection principle. For this example the strategy became too "expensive". Combined with the criteria previously described, we did not reach a final model before extremely decreased popularity on the local network put an end to the experiments.

Anyhow, a final selected model – "ideal" in accordance with a chosen criterion or not – should be judged on how plausible its interpretation seems to the investigator(s). The computational extremely "cheap" strategies implemented in BIFROST combined with an ability to try out and adjust models in the expert system shell HUGIN allow for a great possibility of experimentation.

Finally, it is possible in BIFROST to impose expert(s) a priori knowledge concerning presence and absence of certain interactions within a model by defining a minimal and a maximal model. This is done by specifying edges of the graphical representation which must be included and/or excluded. Besides the impact on the quality of a final selected model, this type of a priori knowledge also has a considerably reducing effect on the set of selectable models.

## 4. The expert system

As described in Section 2 and illustrated in Figure 1, a final block recursive model is selected by applying the chosen selection method to the set of graphical models associated with each block and then combining the final individual selected models into the block recursive model. By only considering decomposable graphical models the selected block recursive model is decomposable.

In itself, the final selected block recursive model only reflects logical (in) dependencies. However, specific estimates of probabilities are required to initialise a probabilistic expert system.

The current release of the HUGIN expert system shell operates on quantified recursive models, also denoted as causal probabilistic networks. Therefore, the export to HUGIN of a quantified final model depends on the fact that any decomposable block recursive model can be represented by an equivalent recursive model.

A complete ordering $\prec$ of the nodes is *compatible*, if

$$\alpha < \beta \Rightarrow \alpha \prec \beta,$$

were $<$ refers to the partial ordering given by the chain graph of the block recursive model. A compatible ordering is said to be *locally reducible* if $\rho \in \Delta(t)$ and $\alpha, \beta \in pa(\rho)$ w.r.t. the ordering $\prec$ implies

$$\alpha \in pa(\beta) \quad \text{or} \quad \beta \in pa(\alpha) \quad \text{or} \quad \alpha, \beta \in \Delta(t).$$

An equivalent recursive model is obtained by finding a compatible and locally reducible ordering of the nodes, and then directing all undirected edges from lower numbered nodes to higher numbered nodes. It is an easy induction argument to show that if a block recursive model is decomposable then there exists an ordering which is compatible and locally reducible.

Now, according to proposition 8.2 of Lauritzen and Wermuth (1989a, 1989b) we have that if a complete ordering $\prec$ is locally reducible and compatible, then the block recursive model and the recursive model ordered by $\prec$ are equivalent. Hence, by a simple calculation of the maximum likelihood estimate (see

e.g. Thiesson, 1991) a quantified recursive model equivalent to the quantified final selected block recursive model is obtained and can then immediately be exported to HUGIN for experimentation.

## 5. The coronary artery disease example

An evaluation of BIFROST is based on a database of observations on variables relevant to coronary artery disease (CAD) (see Hansen, 1980).

The variables are the following: The *Coronary artery disease* (**c**), abbreviated CAD, which we are interested in making diagnoses about, is a disease caused by a reduction in the ability of the coronary arteries to supply the heart muscle. The clinical diagnosis on the state of the disease is made on the basis of coronary arteriography. As background variables there are the *sex* (**s**) of the patient and the risk factors *smoking* (**S**), *hypercholesterolaemia* (**H**), and *hereditary predispositions* (**I**). In addition there is a variable *workload* (**w**), indicating whether the patient had a sufficient heart frequency under the ECG-examinations. This is a composite variable and describes whether the heart frequency plus the age of the patient exceeds 180. As disease manifestations there are *previous myocardial infarct* (**A**), *angina pectoris* (**a**), *hypertrophi* (**h**), and other *heartfailures* (**K**). Finally there are results on the clinical ECG-examinations, that is, information on *Q-wave* (**Q**) and *T-wave* (**T**). Additionally there are two variables *Q-wave informative* (**q**) and *T-wave informative* (**t**) indicating how much confidence to have on the Q-wave and T-wave results. *Angina pectoris* has 3 levels and the remaining 13 variables are binary.

Two data sets are involved: (a) On 236 patients information is available on all 14 variables. Of these patients 107 actually had the disease. This data set is referred to as the *learning* cases. (b) On 67 patients incomplete information is available. That is, information is missing on one or more of the variables *hereditary predispositions, smoking*, and *hypercholesterolaemia*. Of these patients 26 had the disease. This data set is referred to as the *test* cases.

As discussed in Section 1 causal relationships between (some of) the variables suggest that they should not be treated on equal footing. The variables should be partitioned into blocks in accordance with their causal relationship. Furthermore, prior knowledge concerning relations, causal and symmetric, which a priori are known to be present respectively absent should be taken into account. This situation, also discussed in Section 1, is handled by fixing respectively excluding certain edges and arrows presequential to the model selection process, which leads to a specification of a minimal respectively a maximal model.

The characteristics of the variables suggest a rough block recursive structure with 4 blocks. These blocks contain, from the left to the right, the background variables (*sex* and the risk factors), the disease, the disease manifestations, and the clinical results. On the basis of a discussion with a clinician this was refined futher to a block recursive structure with 7 blocks. Furthermore a priori present and absent relations were specified. This is illustrated in Figure 2.

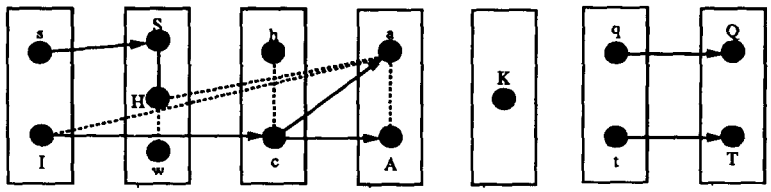



Fig. 2. The block recursive structure obtained from a discussion with the clinician. Relations a priori known to be present are drawn as solid lines, whereas those a priori known to be absent are dashed.

Placing *Hypertrophi* and CAD in the same block may seem strange as *hypertrophi* was characterized as a disease manifestation, whereby it should be in a block to the right of CAD. However *hypertrophi*, which indicates an enlargement of the myocardium, can, according to the clinician, in some cases be an (indirect) consequence of CAD and in other cases it can (indirectly) cause CAD. Being unable to distinguish between these two situations we have chosen to place the variables in the same block.

## 6. Experiments

The purpose of the following experiments is to give a preliminary evaluation of different models selected by BIFROST by investigating the ability of the models to give reliable predictions. A detailed evaluation of the methods and principles on which BIFROST is based is the subject of a prospective paper.

All models selected are based on the block recursive structure and the minimal and maximal models shown in Figure 2. The coherent-direct selection strategy is used for the selection of all models. As selection criterion, IC with different values of the penalty parameter $\kappa$, is used. Thereby models with a wide difference in complexity but not in their basic structure are obtained. Thus the focus will be on determining values of $\kappa$ giving models which, for this particular data set, give reliable predictions.

### *6.1. Evaluation methods*

The evaluation is performed by reclassifying the 236 learning cases and by classifying the 67 test cases. Each selected block recursive model is exported to HUGIN and the probability of CAD given the information available on the remaining variables,

$$P_{\kappa}(CAD \mid evidence),$$

is computed for each case. This probability is compared with the clinical diagnosis based on coronary arteriography by finding the relative frequency of patients actually having CAD in a given interval of predicted probabilities (upper part of Table 1).

A more incisive comparison of the models is made by calibration plots (Figure 5). For each model, the midpoints of the intervals of predicted probabilities are plotted against the relative frequency of CAD in each interval. The closer the points are to the identity line, the better the predictions are.

As a measure of the overall performance of the models we have used the Brier score function, which was first introduced in Brier (1950) as a measure of goodness in the field of probabilistic weather forecasting. By letting $n$ be the number of patients and $p_\kappa = P_\kappa(CAD \mid evidence)$, the Brier score becomes:

$$B_\kappa = \frac{1}{n}\left\{ \sum_{\text{Patients not having CAD}} p_\kappa^2 + \sum_{\text{Patients having CAD}} (1-p_\kappa)^2 \right\}.$$

A good precision in the predictions from a model is reflected by a low value of the score function. For further details on score functions the reader is referred to e.g. Dawid (1986).

The coronary artery disease example has previously been investigated in Andersen, Krebs and Damgaard (1991) who were concerned with semi-automatic selection of an expert system based on a causal probabilistic network. To allow a rough comparison to their results a simpler measure of performance is used. With a decision threshold set on 50%, the relative frequency of misclassifications is found for each system. This measure is, however, a bit unfair to any probabilistic system since it does not take into account the certainty of the diagnoses.

### 6.2. Results

By use of the selection method composed by the coherent-direct selection strategy and *IC* with different values of the penalty parameter $\kappa$ a number of models have been selected. The model selected with $\kappa = 2$, which, as discussed in Section 3, corresponds to *AIC*, is almost saturated, in the sense that only very few edges are removed. By increasing the value of $\kappa$ the simpler the selected models become. Even $\kappa = 3$ leads to the selection of a reasonably simple model. Using $\kappa > 14$ results in models represented by graphs which are not connected. The models obtained for $\kappa = 5$ and $\kappa = 6$ are identical and so are the models for $\kappa = 10, \ldots, 14$.

Note that choosing $\kappa$ between 5 and 6 as penalty parameter is in this example essentially equivalent to using the Bayesian Information Criterion BIC (see Schwarz, 1978) as *BIC* is just the *IC* criterion with penalty parameter $\kappa = \ln$ number of observations $= \ln 236 = 5.46$.

In the following we we will pay particular attention to models obtained by using $\kappa = 2$ (a very complex model), $\kappa = 10$ (a fairly simple model), and $\kappa = 6$ (a value of $\kappa$ in the middle of the two extremes). The HUGIN representation of the model obtained with $\kappa = 6$ is shown in Figure 3.

The results of the (re)classification of the two data sets for these three models are shown in Table 1. In the upper part of the table, the left most column gives, in percentage, the predicted probability of CAD given the available evidence.

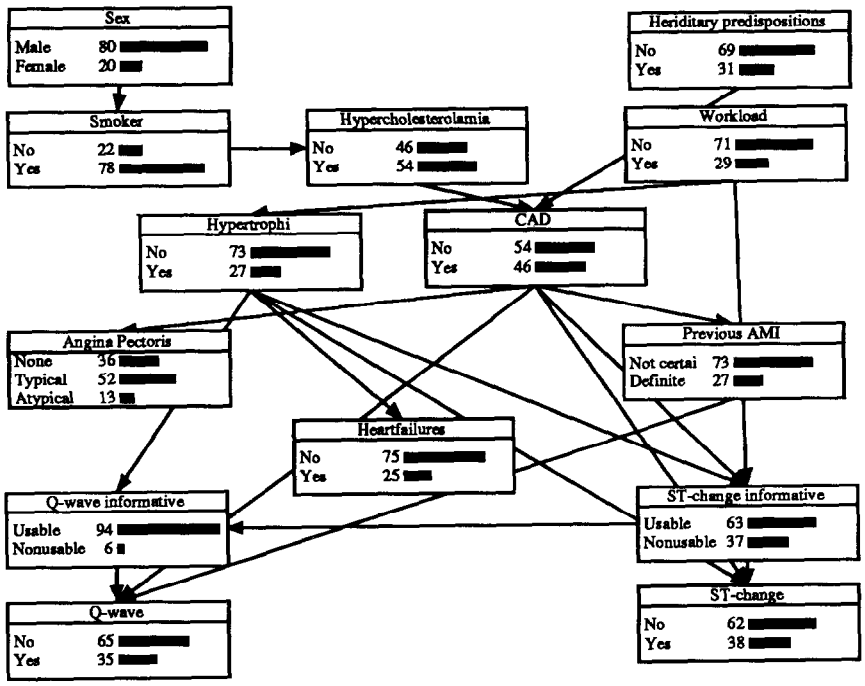



Fig. 3. The HUGIN representation of the block recursive model obtained by using the coherent-direct selection strategy and the IC criterion with penalty parameter $\kappa = 6$. Notice, that for this representation the block recursive model has been converted into an equivalent recursive model. Direction of causality is top-down instead of from the left to the right. The length of the horizontal bars indicate the marginal probabilities of the possible levels of each variable, obtained from the selected fitted model. The numbers in front of the bars represent the same probabilities in percentage.

For each of the remaining columns the denominator gives the number of patients assigned a probability of CAD in the corresponding interval, and the numerator gives the number of these which actually had CAD. Taking 50% as the decision threshold the percentage of falsely negative and falsely positive diagnoses is for each model shown in the middle part of the table. In the lower part the value of the Brier score function is shown for each model.

Consider the classification results on the learning cases. It is to be expected that the error rates are relatively small as we classify the observations on which the models are based. Also, it is not surprising that the error rates as well as the value of the score function in the learning cases increase with $\kappa$ as the number of parameters in the model decrease. Comparing the error rates between the different data sets one would expect these to be larger for the test cases than for the learning cases which is also what is observed. Besides the argument that the models are selected on the basis of the learning cases, the error rate in favor of the learning cases is also a consequence of the fact that the test data has sporadically missing information on *smoking, hypercholesterolamia*, and *hereditary predispositions*. For all models selected there are direct dependencies

Table 1
The classification results for models obtained by using the selection method composed by the coherent-direct strategy and the information criterion with penalty parameters $\kappa = 2$, $\kappa = 6$, and $\kappa = 10$. The upper part: The left most column gives, in percentage, the predicted probability of CAD given the available evidence. In the remaining columns the fraction of patients actually having CAD within the interval of assigned probabilities is shown. The middle part: Taking 50% as the decision threshold the percentage of falsely negative and falsely positive is shown for each model. The lower part: The value of the Brier score function for each model.

| P(CAD \| evidence) (%) | The learning cases (236) | | | The test cases (67) | | |
|---|---|---|---|---|---|---|
| | $\kappa = 2$ | $\kappa = 6$ | $\kappa = 10$ | $\kappa = 2$ | $\kappa = 6$ | $\kappa = 10$ |
| [0;1[ | 0/71 | 0/27 | 0/11 | 2/10 | 0/4 | 0/1 |
| [1;10[ | 0/29 | 1/52 | 0/57 | 4/16 | 3/24 | 1/19 |
| [10;30[ | 3/21 | 3/30 | 7/37 | 3/5 | 4/10 | 6/11 |
| [30;50[ | 1/7 | 12/26 | 7/21 | 3/12 | 4/10 | 3/11 |
| [50;70[ | 11/13 | 10/15 | 17/29 | 1/2 | 3/5 | 5/9 |
| [70;90[ | 7/8 | 17/19 | 29/31 | 1/3 | 7/8 | 6/10 |
| [90;99[ | 10/12 | 47/50 | 38/41 | 2/2 | 5/6 | 5/6 |
| [99;100] | 75/75 | 17/17 | 9/9 | 10/17 | 0/0 | 0/0 |
| Error rates (%) | | | | | | |
| Falsely negative | 3.1 | 11.9 | 11.1 | 27.9 | 22.9 | 23.8 |
| Falsely positive | 4.6 | 9.9 | 15.5 | 41.7 | 21.1 | 36.0 |
| Brier score ($\times 100$) | 37 | 79 | 96 | 288 | 167 | 193 |

between *hypercholesterolamia* and CAD respectively *hereditary predispositions* and CAD. This means that valuable information may be missing.

The missing information also has the interesting consequence, that the results on the test cases are much less affirmative than for the learning data. That is, a larger fraction of the patients are assigned a probability in the middle of the interval, and fewer at the end points compared to the learning data.

Naturally, the quality of the models should be judged on their performance on the test cases, as this reflects the predictive ability on future cases. In Table 1 it can be seen that for the test cases the error rates as well as score values are smallest for $\kappa = 6$. In Figure 4 the values of the Brier score function is plotted against different values of $\kappa$, for the learning as well as the test cases. The plot

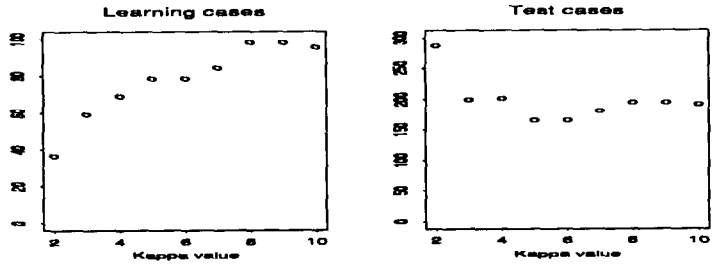



Fig. 4. The Brier scores ($\times 1000$) for models obtained by the coherent-direct strategy and the information criterion are on the y-axis plotted against the corresponding values of the penalty parameter $\kappa$. The left most plot is for the learning cases and the right most is for the test cases.

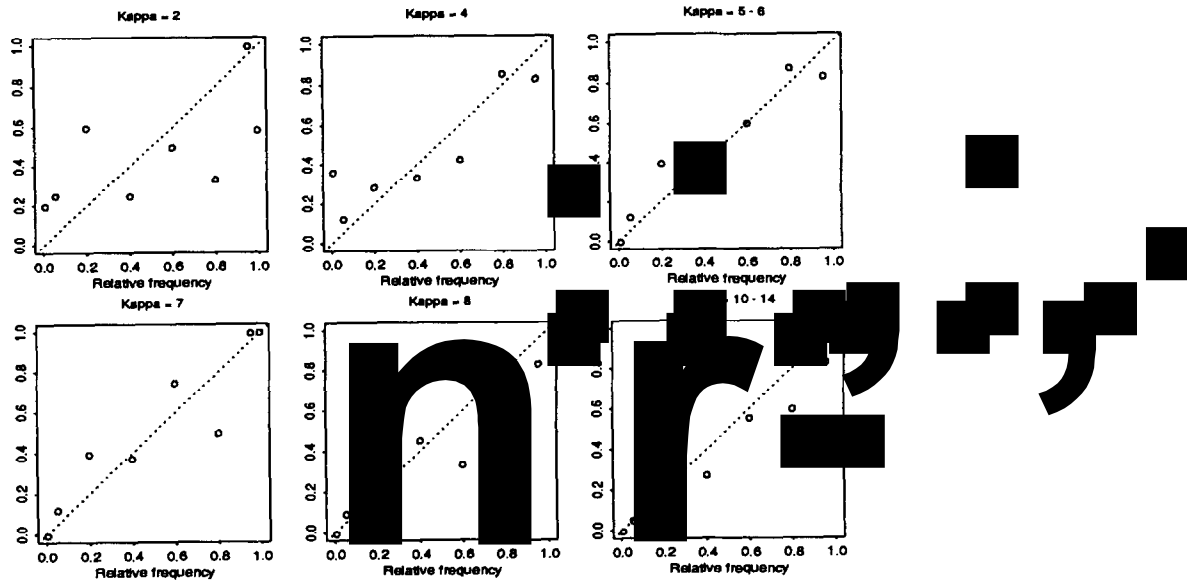



Fig. 5. Calibration plots for models obtained by the coherent-direct strategy and the information criterion with different values of the penalty parameter $\kappa$. On the y-axis, the midpoints of the intervals of predicted probabilities (given in Table 1) are plotted against the relative frequency of patients in the test cases having CAD in this interval.

for the test cases indicates that $\kappa = 6$ gives a model with good overall performance. Comparing the two plots in Figure 4, it is interesting to see that for $\kappa$-values larger than or equal 5 the shapes of the plots are alike.

A more incisive investigation of the performance of the selected models is made by the calibration plots in Figure 5. For $\kappa = 6$ it appears that the fraction of patients actually having CAD among the patients assigned to an interval of predicted probabilities for the disease is in good accordance with the assigned probabilities for having the disease.

To summarize, the best results are in this example achieved when $\kappa$ is chosen in the middle of the interval between the highest value of $\kappa$ for which the selected model for all practical purposes is saturated and the smallest value of $\kappa$ which result in the most simple model represented by a connected graph. For $\kappa$ equal 6 and with the very rough decision threshold on 50%, the total probability of a wrong diagnosis can be extracted from Table 1 as

$$P(\text{Wrong diagnosis}) = P(\text{Falsely negative})P(\text{Having CAD}) + P(\text{Falsely positive})P(\text{Not having CAD}) \approx 0.22.$$

Thus the overall risk for an erroneous diagnosis is 22%, which seems to be the best possible result in this example. This result may be compared to an overall risk for an erroneous diagnosis on 24% for the system reported in Andersen et al. (1991). The slightly better performance of the system selected by BIFROST may have been caused by peculiarities in the data set and by the fact that a

direct comparison was impossible due to a different relation between learning and test cases. So, with respect to this measure, it is not clear, which of the two systems would perform better in future cases. It is possible, however, that a measure more fair to the probabilistic nature of the diagnosis obtained by the two systems may unveil differences.

## 7. Acknowledgements

The authors wish to acknowledge Jørgen Greve and Flemming Skjøth who participated in the original BIFROST project, which motivated this article. Also, we wish to thank Steffen L. Lauritzen for inspiring the creation of BIFROST. The BIFROST program is implemented as a very high level application upon the program CoCo (Badsberg, 1991). Hence, most computational features are obtained by utilizing access to this program. We wish to acknowledge Jens Henrik Badsberg for allowing this access and for friendly assistance with CoCo. We thank Jørgen Fischer Hansen for helpful conversations and for allowing us to use the coronary artery disease data material. Finally we thank the referees for their helpful suggestions and Harri T. Kiiveri for proof reading the final manuscript.

## References

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Transactions on Automatic Crontrol* **19**(6) 716–723.

Andersen, L.R., Krebs, J.H. and Damgaard, J. (1991), STENO: an expert system for medical diagnosis based on graphical models and model search, *Journal of Applied Statistics*.

Andersen, S.K., Olesen, K.G., Jensen, F.V. and Jensen, F. (1989), HUGIN – a shell for building Bayesian belief universes for expert systems, *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (IJCAI)*, Detroit, MI.

Badsberg, J.H. (1991), A guide to CoCo, *Technical Report R 91-43*, Institute for Electronic Systems, Aalborg University.

Brier, G.W. (1950), Verification of forecasts expressed in terms of probability, *Monthly Weather Review* **78**(1) 1–6.

Darroch, J.N., Lauritzen, S.L. and Speed, T.P. (1980), Markov fields and log linear models for contingency tables, *Annals of Statistics* **8** 522–539.

Dawid, A.P. (1986), Probability forecasting, *in* S. Kotz and N.L. Johnson (eds), *Encyclopedia of statistical sciences*, Vol. 7, Wiley.

Edwards, D. (1984), A computer intensive approach to the analysis of sparse multidimensional contingency tables, *COMPSTAT*, Physica-Verlag, Vienna for IASC, pp. 355–359.

Edwards, D. and Havránek, T. (1985a), A fast procedure for model search in multidimensional contingency tables, *Biometrika* **72** 339–351.

Edwards, D. and Havránek, T. (1985b), A fast procedure for model search in multidimensional contingency tables, *Biometrika* **72**(2) 339–351.

Edwards, D. and Havránek, T. (1987a), A fast model selection procedure for large families of models, *Journal of the American Statistical Association* **82** 205–211.

Edwards, D. and Havránek, T. (1987b), A fast model selection procedure for large families of models, *Journal of the American Statistical Association* **82**(397) 205–211.

Fischer, L.P. (1990), *HUGIN REGULAR User's Guide*, version 1.0 edn, HUGIN Expert Ltd., Denmark.

Frydenberg, M. (1990a), The chain graph Markov property, *Scandinavian Journal of Statistics* **17** 333–853.

Frydenberg, M. (1990b), Marginalization and collapsibility in graphical interaction models, *Annals of Statistics* **18** 790–805.

Frydenberg, M. and Lauritzen, S.L. (1989), Decomposition of maximum likelihood in mixed interaction models, *Biometrika* **76** 539–555.

Gabriel, K.R. (1969), Simultaneous test procedures – some theory of multiple comparisons, *Annals of Mathematical Statistics* **40** (1) 224–250.

Goodman, L.A. (1971), The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications, *Technometrics* **13** 33–61.

Greve, J., Højsgaard, S., Skjøth, F. and Thiesson, B. (1990), Automatic model selection in contingency tables, Institute for Electronic Systems, Aalborg University.

Hansen, J.F. (1980), The clinical diagnosis of ischaemc heart disease due to coronary artery disease, *Danish Medical Bulletin*.

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

Havránek, T. (1990), On application of statistical model search techniques in constructing a probabilistic knowledge base, In: Trans. of the Eleventh Prague Conference, 375–377.

Højsgaard, S., Skjøth, F. and Thiesson, B. (1992), User's guide to BIFROST, *Technical Report R-92-2001*, Institute for Electronic Systems, Aalborg University.

Lauritzen, S.L. (1989a), Lectures on contingency tables, *Technical Report R-89-29*, Institute for Electronic Systems, Aalborg University. (3rd. edn).

Lauritzen, S.L. (1989b), Mixed graphical association models (with discussion), *Scandinavian Journal of Statistics* **16** 273–306.

Lauritzen, S.L. (1992), Propagation of probabilities, means and variances in mixed graphical association models, *Journal of the American Statistical Association* **87** 1098–1108.

Lauritzen, S.L. and Wermuth, N. (1984), Mixed interaction models, *Technical Report R 84-8*, Institute for Electronic Systems, Aalborg University.

Lauritzen, S.L. and Wermuth, N. (1989a), Correction to (Lauritzen and Wermuth 1989b), *Annals of Statistics* **17**(4) 1916.

Lauritzen, S.L. and Wermuth, N. (1989b), Graphical models for associations between variables, some of which are qualitative and some quantitative, *Annals of Statistics* **17** 31–57.

Lauritzen, S.L., Dawid, A.P., Larsen, B.N. and Leimer, H.-G. (1990), Independence properties of directed Markov fields, *Networks* **20** 491–505.

Lauritzen, S.L., Speed, T.P. and Vijayan, K. (1984), Decomposable graphs and hypergraphs, *Journal of the Australian Mathematical Society, Series A* **36** 12–29.

Linhart, H. and Zucchini, W. (1986), *Model Selection*, Wiley, New York.

Madigan, D. and Raftery, A.E. (1991), Model selection and accounting for model uncertainty in graphical models using occam's window, *Technical report*, Department of Statistics, GN-22, University of Washington.

Patefield, W.M. (1981), Algorithm AS 159. An efficient method of generating random $r \times c$ tables with given row and column totals, *Applied Statistics* **30** 91–97.

Schwarz, G. (1978), Estimating the dimension of a model, *Annals of Statistics* **6**(2) 461–464.

Thiesson, B. (1991), *(G)EM algorithms for recursive graphical association models with missing data*, Master's thesis, Institute for Electronic Systems, Aalborg University.

Wermuth, N. (1976), Model search among multiplicative models, *Biometrics* **32** 253–263.

Wermuth, N. and Lauritzen, S.L. (1983), Graphical and recursive models for contingency tables, *Biometrika* **70** 537–552.

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*, Wiley.