The Annals of Statistics
1989, Vol. 17, No. 1, 31–57

GRAPHICAL MODELS FOR ASSOCIATIONS BETWEEN VARIABLES, SOME OF WHICH ARE QUALITATIVE AND SOME QUANTITATIVE

BY S. L. LAURITZEN AND N. WERMUTH

Aalborg University and University of Mainz

We define and investigate classes of statistical models for the analysis of associations between variables, some of which are qualitative and some quantitative. In the cases where only one kind of variables is present, the models are well-known models for either contingency tables or covariance structures. We characterize the subclass of decomposable models where the statistical theory is especially simple. All models can be represented by a graph with one vertex for each variable. The vertices are possibly connected with arrows or lines corresponding to directional or symmetric associations being present. Pairs of vertices that are not connected are conditionally independent given some of the remaining variables according to specific rules.

1. Introduction. The purpose of the present article is to develop statistical models, with discrete and continuous random variables, that can be used to describe and investigate associations among properties of observational units, some of which are qualitative and some of which are quantitative.

The applications we have in mind are primarily in the social sciences and we believe that it is indispensable that the models can take into account that variables can be explanatory, responses and both, in the sense that they are responses to some variables but then explanatory for others. The associations between other variables might appropriately be interpreted without this distinction because the variables appear on a symmetric footing.

As an illustration, consider the following example, taken from a thesis of Schumann (1986), referring to cognitive developments in young children.

We shall not go into details, but in the experiment each of 55 children aged from four to seven was confronted with 19 similar tasks, two of which were extremely easy and only included to keep the child motivated.

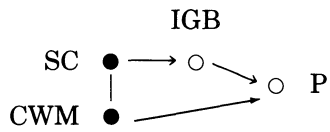
In each task the child had to compare one picture, called the standard, to six others arranged in a row 12 cm apart from the standard. Five of the alternatives differed from the standard in one to at most five characteristics. The total number of correct matches in the 17 tasks was taken as a measure of the *performance* (P), the total number of times the child looked from the standard to the alternatives or back (divided by 17) as a measure for the *information gathering behaviour* (IGB). Furthermore, the *capacity of the working memory* (CWM) was operationalized by the 'digit span backwards,' that is, by the number of digits the child could repeat in reverse order, while the pure *storage*

Received May 1987; revised June 1988.

AMS 1980 *subject classifications.* Primary 62H99; secondary 62J99.

Key words and phrases. Analysis of variance, conditional independence, contingency tables, covariance selection, exponential families, logistic regression, log-linear models, Markov random fields, multivariate analysis, path analysis, regression, recursive models, triangulated graphs.

capacity (SC) was captured by the 'digit span forwards,' that is, by the number of digits the child could repeat in unchanged order. The levels of SC are called low for less than four digits and high otherwise, while the levels of CWM are low (< 2), medium ($= 2$) and high (> 2). Since the first two variables are viewed as being quantitative, the other as qualitative, the variables P and IGB are analysed as continuous and CWM and SC as discrete. The hypotheses considered by Schumann are summarized in the following graph:



The storage capacity (SC) and the capacity of the working memory (CWM) are explanatory variables on equal footing, the information gathering behaviour (IGB) is conceived as a response to these and the performance as a response to all of them. The (vague) meaning of the picture is that CWM has no influence on IGB, other than what is explainable through SC and that the performance P depends directly only upon CWM and IGB.

One aim of the present article is to enable the researcher to give a *precise* meaning to such pictures by developing corresponding statistical models. The models are appropriate for what Holland (1986) terms 'associational inference.' They are extensions to models known as path analysis [Wright (1921, 1923, 1934)] and provide an alternative to other similar developments in social sciences; see, for example, Wold (1954), Simon (1957), Blalock (1971), Jöreskog (1977), Goodman (1973) and Goldberger and Duncan (1978). See Kiiveri and Speed (1982) and Wermuth (1985) for further discussion and references.

The most general models discussed here are the *graphical chain models* (Section 8). These have previously been discussed in the discrete case by Goodman (1973), Asmussen and Edwards (1983) and in the continuous case by Porteous (1985b).

Special cases of these are the *recursive models* (Sections 5 and 6) investigated in the discrete case by Wermuth and Lauritzen (1983), the continuous case by Wermuth (1980) and Kiiveri (1983) and some aspects of the general case by Kiiveri (1983) and Kiiveri, Speed and Carlin (1984). The latter reference also deals with aspects of the graphical chain models in the general case. Other special cases are the class of graphical Markov models (Sections 3 and 4) that in the discrete case specialize to those of Darroch, Lauritzen and Speed (1980) and in the continuous case to the covariance selection models of Dempster (1972). These models, where all associations are symmetric, are the basic building blocks for the other models. Consequently, most of the article (Sections 3 and 4) is devoted to a study of these and the corresponding distributions. The multivariate distributions upon which our developments are based, are characterized by the joint *conditional* distribution of the continuous variables, given the discrete as being *Gaussian* and therefore called CG-distributions. It is crucial to get a good understanding of these and their interplay with properties of conditional

independence (Sections 2 and 3). For the latter notion as well as our notation on this point, the reader is referred to Dawid (1979, 1980).

The class of models where the statistical theory and interpretation is especially simple has been identified as the *decomposable* models (Section 7). Some technical matters such as our graphtheoretic terminology and a proof of an important result are deferred to the appendices. We suggest that these are omitted at first reading.

Our emphasis here is on the formal development of the models and their properties. For examples of their application and a discussion of some implications of the results for practical statistical work, see Edwards (1987, 1988, 1989) and Wermuth and Lauritzen (1989).

2. CG-distributions and CG-regressions. The present section is devoted to the study of the class of multivariate distributions upon which the models are based. We consider a finite set V of variables partitioned into discrete and continuous as $V = \Delta \cup \Gamma$. Let $|V| = p + q$, $|\Delta| = p$ and $|\Gamma| = q$. Thus our random variables take values in the product space

$$\mathscr{X} = \mathscr{I} \times \mathscr{Y} = \bigtimes_{\alpha \in V} \mathscr{X}_\alpha$$

with $\mathscr{I} = \bigtimes_{\delta \in \Delta} \mathscr{I}_\delta$, $\mathscr{Y} = \mathbb{R}^\Gamma$, where $\mathscr{I}_\delta$, $\delta \in \Delta$, are finite sets of possible levels of the discrete variables.

The corresponding random variables shall be denoted X_α, $\alpha \in \Delta \cup \Gamma$. Thus the variables X_δ, $\delta \in \Delta$, are discrete valued (ranging in $\mathscr{I}_\delta$), whereas X_γ, $\gamma \in \Gamma$, are real-valued.

Typical points of $\mathscr{X}$ are denoted by x or as $x = (i, y)$. Similarly, i_a, y_b, x_d, etc., are used to denote the projections of a point $x = (i, y)$ onto the spaces

$$\mathscr{I}_a = \bigtimes_{\delta \in a} \mathscr{I}_\delta, \qquad \mathscr{Y}_b = \mathbb{R}^b, \qquad \mathscr{X}_d = \bigtimes_{\alpha \in d} \mathscr{X}_\alpha, \quad \text{respectively.}$$

Analogously, we use the notation X_a for the collection of variables $(X_\alpha,\ \alpha \in a)$, and the short notation $a \perp\!\!\!\perp b \mid c$ to indicate that the random variables X_a and X_b are conditionally independent given X_c.

Our investigations shall be directed toward a special class of probability distributions that all have strictly positive density f (w.r.t. product of counting measure on $\mathscr{I}$ and Lebesgue measure on $\mathscr{Y}$) of the form

$$\begin{aligned} f(x) &= f(i, y) \\ &= \exp\left\{g(x_\Delta) + h(x_\Delta)^{\mathrm{T}} x_\Gamma - \tfrac{1}{2} x_\Gamma^{\mathrm{T}} K(x_\Delta) x_\Gamma\right\} \\ &= \exp\left\{g(i) + h(i)^{\mathrm{T}} y - \tfrac{1}{2} y^{\mathrm{T}} K(i) y\right\}, \end{aligned} \tag{2.1}$$

where g is a real-valued function of i, h is a q-vector-valued function of i taking values in $\mathbb{R}^\Gamma$, K is a $q \times q$ matrix-valued function of i taking values in the set of positive definite symmetric matrices and v^{T} denotes the transpose of the vector v.

A probability distribution with density of the form (2.1) has *conditional Gaussian distributions* in the sense that X_Γ for given $X_\Delta = i$ is q-variate Gaussian with covariance $K(i)^{-1}$ and expectation $K(i)^{-1}h(i)$, that is,

$$\mathscr{L}(X_\Gamma \mid X_\Delta = i) = \mathscr{N}_q\big(K(i)^{-1}h(i), K(i)^{-1}\big). \tag{2.2}$$

The marginal distribution of the discrete variables X_Δ has probabilities equal to

$$p(i) = (2\pi)^{-q/2} \det\big(K(i)\big)^{-1/2} \exp\Big\{g(i) + \tfrac{1}{2}h(i)^{\mathrm{T}}K(i)^{-1}h(i)\Big\}. \tag{2.3}$$

To derive these facts, we first rewrite (2.1) as

$$f(x) = \exp\Big\{g^*(i) - \tfrac{1}{2}\big(y - \xi(i)\big)^{\mathrm{T}}K(i)\big(y - \xi(i)\big)\Big\},$$

where we have let

$$\xi(i) = K(i)^{-1}h(i), \qquad g^*(i) = g(i) + \tfrac{1}{2}h(i)^{\mathrm{T}}K(i)^{-1}h(i).$$

We then integrate over y and obtain (2.3) and (2.2).

On the other hand, it is clear that any probability distribution on $\mathscr{I} \times \mathscr{Y}$ with strictly positive marginal probability on $\mathscr{I}$ and with conditional distributions of the continuous variables being multivariate regular Gaussian, with expectation $\xi(i)$ and covariance matrix $\Sigma(i)$, will have an expansion as (2.1), where

$$K(i) = \Sigma(i)^{-1}, \qquad h(i) = K(i)\xi(i),$$

$$g(i) = \log p(i) - \frac{q}{2}\log(2\pi) + \frac{1}{2}\log\det K(i) - \frac{1}{2}h(i)^{\mathrm{T}}K(i)^{-1}h(i).$$

Distributions of this type shall be called *CG-distributions* and are the basic distributions entering in the present article. It is of interest also to consider the special case when the covariance matrix of the conditional distribution of the continuous variables given the discrete ones does not depend on i, that is $K(i) \equiv K$, in which case we shall say that the probability distribution is *homogeneous* and call it an HCG-distribution.

The latter class of distributions (HCG) was considered by Tate (1954) and Olkin and Tate (1961) in a context of defining correlation among a binary and a continuous variable, and by Dempster (1973) in studying aspects of a so-called multinomial logit model. Later, the distributions have been used by Krzanowski (1983) and Little and Schluchter (1985) in a context different from the present. Since we shall be interested in HCG-distributions as well as CG-distributions we shall adopt the convention that *unless otherwise stated, all theorems, statements, etc., about CG-distributions, remain true if CG everywhere is replaced by HCG*. In cases where this is not obvious, we shall comment on that explicitly.

A CG-distribution can be specified either by the triple (g, h, K) or (p, ξ, Σ), whichever might be convenient in the context considered. We shall term the first triple as the *canonical* and the second as the *moment characteristics* of the CG-distributions.

An important key to understanding the development in the following sections is the behaviour of these distributions under *conditioning* and *marginalization*.

Let $V = A \cup B$ be a partitioning of the set of variables. We then have the following proposition.

PROPOSITION 2.1. *If $B \subseteq \Gamma$ and X has a CG-distribution, the marginal distribution of X_A is CG.*

PROOF. The result follows immediately by integrating (2.1) over x_B $(= y_B)$. □

In general this is not true for $B \subseteq \Delta$. However,

PROPOSITION 2.2. *If X has a CG-distribution and $B \subseteq \Delta$ satisfies*

$$B \perp\!\!\!\perp \Gamma \mid \Delta \setminus B,$$

then the marginal distribution of X_A is CG.

PROOF. By standard properties of conditional independence, the condition ensures that

$$\mathscr{L}(X_\Gamma \mid X_{\Delta \setminus B}) = \mathscr{L}(X_\Gamma \mid X_\Delta),$$

and the latter is Gaussian by assumption. □

For conditional distributions we have the following proposition.

PROPOSITION 2.3. *If X has a CG-distribution, the conditional distribution of X_A given $X_B = x_B$ is CG.*

PROOF. The result follows from the identity

$$\mathscr{L}(X_{A\cap\Gamma} \mid X_{A\cap\Delta} = i_{A\cap\Delta}, X_B = x_B) = \mathscr{L}(X_{A\cap\Gamma} \mid X_\Delta = i_\Delta, X_{B\cap\Gamma} = x_{B\cap\Gamma})$$

and the latter is Gaussian since it is obtained by conditioning upon $x_{B\cap\Gamma}$ in the (conditional) Gaussian distribution of X_Γ, given $X_\Delta = i_\Delta$. □

For later use, we need to expand upon Proposition 2.3 and consider *how* this conditional distribution depends upon x_B. We thus introduce the class of *CG-regressions* as systems of maps that to any element (j, z) of a product set

$$\mathscr{I} \times \mathscr{Z} = \left(\bigtimes_{\delta \in \Delta^*} \mathscr{I}_\delta \right) \times \mathbb{R}^{\Gamma^*}$$

of possible state spaces for discrete and continuous variables, specify a CG-distribution on $\mathscr{I} \times \mathscr{Y}$ with moment characteristics (p, ξ, Σ) depending on (j, z) in a particular way:

$$\begin{aligned} \log p(i \mid j, z) &= u(i \mid j) + v(i \mid j)^{\mathrm{T}} z + z^{\mathrm{T}} W(i \mid j) z - \log \kappa(j, z), \\ \xi(i \mid j, z) &= a(i \mid j) + B(i \mid j) z, \\ \Sigma(i \mid j, z) &= C(i \mid j). \end{aligned} \tag{2.4}$$

The CG-regression is specified by the sixtuple (u, v, W, a, B, C). The term $\kappa(j, z)$ is a normalizing constant,

$$\kappa(j, z) = \sum_i \exp\{u(i \mid j) + v(i \mid j)^{\mathrm{T}} z + z^{\mathrm{T}} W(i \mid j) z\},$$

depending on (u, v, W). If we change any of (u, v, W) by adding terms depending on j only, they will eventually cancel out in the final expression (2.4), giving rise to the same $p(i \mid j, z)$.

A CG-regression specifies a *quadratic dependence* of $\log p$ on z, a *linear dependence* on z of the conditional expectation and a *nondependence* on z of the conditional covariance matrix. The coefficients of the dependencies as well as the conditional covariance matrix are allowed to depend on j.

We shall say that a CG-regression is *homogeneous* or an *HCG-regression* if $W(i \mid j) \equiv 0$, $B(i \mid j) \equiv B$, $C(i \mid j) \equiv C$, in words if $\log p$ depends *linearly* on z, the linear dependencies of conditional expectations are *parallel* and the covariance matrix is *constant*. We then have the following proposition.

PROPOSITION 2.4. *A sixtuple* (u, v, W, a, B, C) *specifies a CG-regression if and only if there is a joint CG-distribution on* $(\mathcal{I} \times \mathcal{J}) \times (\mathcal{Y} \times \mathcal{Z})$ *such that* (2.4) *specifies the conditional distribution of* (I, Y) *given* $(J, Z) = (j, z)$.

PROOF. Suppose that we have a joint CG-distribution with characteristics (p, ξ, Σ) and let us partition ξ, K and Σ as

$$\xi = \begin{pmatrix} \xi_Y \\ \xi_Z \end{pmatrix}, \qquad K = \begin{pmatrix} K_{YY} & K_{YZ} \\ K_{ZY} & K_{ZZ} \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ} \end{pmatrix}.$$

Straightforward calculations yield that the conditional distributions have characteristics (2.4) given as

$$\begin{aligned} C(i \mid j) &= K_{YY}^{-1}(i, j), \\ B(i \mid j) &= -C(i \mid j) K_{YZ}(i, j), \\ a(i \mid j) &= \xi_Y(i, j) - B(i \mid j) \xi_Z(i, j), \\ W(i \mid j) &= \tfrac{1}{2} \Sigma_{ZZ}^{-1}(i, j), \\ v(i \mid j) &= \Sigma_{ZZ}^{-1}(i, j) \xi_Z(i, j), \\ u(i \mid j) &= \log p(i, j) - \tfrac{1}{2} \log \det \Sigma_{ZZ}(i, j) \\ &\qquad - \tfrac{1}{2} \xi_Z(i, j)^{\mathrm{T}} \Sigma_{ZZ}^{-1}(i, j) \xi_Z(i, j). \end{aligned} \tag{2.5}$$

This direction was the obvious one. To show the converse, we have to fill the apparent gap that W does not have to be positive definite whereas Σ must be, implying that (2.5) cannot be used directly. So suppose conversely that we have given a CG-regression by a sixtuple (u, v, W, a, B, C), and we want to construct a CG-distribution with corresponding conditional distributions. We then first let

$$D(i, j) = \theta E - 2W(i \mid j),$$

where E is the identity matrix, and choose $\theta > 2\,|\lambda_{\max}|$, where $\lambda_{\max}$ is the largest in absolute value of the eigenvalues of all the matrices $W(i\,|\,j)$, thereby obtaining that $D(i, j)$ is positive definite for all (i, j). We then let

$$K_{ZZ}(i, j) = D(i, j) + B(i\,|\,j)^{\mathrm{T}}C(i\,|\,j)^{-1}B(i\,|\,j),$$

$$K_{YZ}(i, j) = -C(i\,|\,j)^{-1}B(i\,|\,j) = K_{ZY}^{\mathrm{T}}(i, j), \qquad K_{YY}(i, j) = C(i\,|\,j)^{-1}.$$

The matrix $K(i, j)$ so defined is then positive definite since for arbitrary $e^{\mathrm{T}} = (y^{\mathrm{T}}, z^{\mathrm{T}}) \neq 0$ we have, suppressing the dependence on (i, j),

$$\begin{aligned} e^{\mathrm{T}}Ke &= y^{\mathrm{T}}C^{-1}y - 2y^{\mathrm{T}}C^{-1}Bz + z^{\mathrm{T}}Dz + z^{\mathrm{T}}BC^{-1}Bz \\ &= (y - Bz)^{\mathrm{T}}C^{-1}(y - Bz) + z^{T}Dz > 0. \end{aligned}$$

We can now determine ξ from a and v by the equations (2.5). Finally, $p(i, j)$ can be calculated. Doing the calculations in reverse order, remembering that W is only determined up to an additive term, we get that our CG-distribution so constructed has the right conditional distributions. □

If the joint distribution is homogeneous we get $B(i\,|\,j) \equiv B$, $C(i\,|\,j) \equiv C$ and $W(i\,|\,j) \equiv W$. Recalling that W is only determined up to an additive term, possibly depending on j, we might as well take $W \equiv 0$, giving a homogeneous regression. On the other hand, had the CG-regression been homogeneous, K, as defined, would not depend on (i, j) and the resulting CG-distribution would therefore be homogeneous.

In the special case where the set of 'response' variables (I, Y) has only one element, $V = \Delta = \{\delta\}$ or $V = \Gamma = \{\gamma\}$, we use the term *univariate CG-regressions*.

As an illustration of some of the previous developments consider the following two examples, to be used throughout the remainder of the article:

1. involving one discrete and two continuous variables (I, Y_1, Y_2) and
2. involving two discrete and one continuous variable (I_1, I_2, Y).

In the first case, the general log density becomes

$$\begin{aligned} \log f(i, y_1, y_2) = {} & g(i) + h^{(1)}(i)y_1 + h^{(2)}(i)y_2 \\ & - \tfrac{1}{2}\left[k^{(11)}(i)y_1^2 + 2k^{(12)}(i)y_1y_2 + k^{(22)}(i)y_2^2\right]. \end{aligned} \tag{2.6}$$

If the distribution is homogeneous, $k^{(11)}$, $k^{(12)}$ and $k^{(22)}$ do not depend on i. The marginal distribution of (I, Y_1) is conditional Gaussian, but this is typically *not* the case for the marginal distribution of (Y_1, Y_2). In the homogeneous case, the conditional distribution of I, given $(Y_1, Y_2) = (y_1, y_2)$ has the form

$$\log p(i\,|\,y_1, y_2) = g(i) + h^{(1)}(i)y_1 + h^{(2)}(i)y_2 - \log \kappa(y_1, y_2).$$

In example 2 the general log density is

$$\log f(i_1, i_2, y) = g(i_1, i_2) + h(i_1, i_2)y - \tfrac{1}{2}k(i_1, i_2)y^2, \tag{2.7}$$

the marginal distributions of (I_1, Y) or (I_2, Y) are not CG in general, whereas the

conditional distribution of Y, given $(I_1, I_2) = (i_1, i_2)$, is Gaussian with expectation $h(i_1, i_2)/k(i_1, i_2)$ and variance $k(i_1, i_2)^{-1}$, the latter being independent of (i_1, i_2) in the homogeneous case.

3. The Markov property for CG-distributions and CG-interactions. Consider the setup in the previous section and let a marked graph $\mathcal{G} = (\Delta \cup \Gamma, E)$ be given. In the present section $\mathcal{G}$ is always assumed to be *undirected*, that is, each connection between vertices is a line.

A distribution on

$$\mathcal{X} = \mathcal{I} \times \mathcal{Y} = \left(\bigtimes_{\delta \in \Delta} \mathcal{I}_\delta \right) \times \mathbb{R}^\Gamma$$

with strictly positive density f is said to be Markovian w.r.t. G or *G-Markovian* if it satisfies

$$\mathbf{M}: \quad \alpha \notin \operatorname{adj}(\beta) \Rightarrow \{\alpha\} \perp\!\!\!\perp \{\beta\} \mid V \setminus \{\alpha, \beta\} \quad \text{for all } \alpha, \beta \in V,$$

that is, if pairs of variables corresponding to nonadjacent vertices are conditionally independent given the remaining variables.

To introduce the notion of interaction, let us reconsider the general expression for a CG-distribution,

$$\log f(i, y) = g(i) + h(i)^{\mathrm{T}} y - \tfrac{1}{2} y^{\mathrm{T}} K(i) y. \tag{3.1}$$

Let us also adopt the convention that functions denoted by f_a, ψ_a, λ_a, etc., depend on $x = (i, y)$ only through its coordinates in a, $(x_\alpha, \alpha \in a) = x_a$, in other words $x_a = z_a \Rightarrow f_a(x) = f_a(z)$, etc. Thus functions $f_\varnothing, \psi_\varnothing$, etc., with the empty set as subscript are constant. We can now make expansions as follows:

$$g(i) = \sum_{d \subseteq \Delta} \lambda_d(i), \qquad h(i) = \sum_{d \subseteq \Delta} \eta_d(i), \qquad K(i) = \sum_{d \subseteq \Delta} \Psi_d(i).$$

In general such expansions can be made in many ways, see, for example, Darroch and Speed (1983) for a comprehensive discussion of this and similar problems. Whenever such an expansion has been made, we shall denote the λ, η and Ψ terms as *interactions* and give them special names:

$\lambda_\varnothing$ is the *log normalizing constant.*

λ_d, $d \neq \varnothing$, are *pure discrete interactions* among variables in d. If $|d| = 1$ we also call these *main effects* of the discrete variables.

$\eta_\varnothing$'s coordinates are the *main effects* of the continuous variables.

η_d, $d \neq \varnothing$, are *mixed linear interactions* and its coordinates are the mixed linear interaction between a continuous variable and variables in d.

Ψ_d, $d \subseteq \Delta$, are *quadratic interaction matrices*; the elements of $\Psi_\varnothing(i)$ do not depend on i and are called *pure quadratic interactions*. The elements of $\Psi_d(i)$, $d \neq \varnothing$, are *mixed quadratic interactions* between variables in d and pairs of continuous variables.

Note that a CG-distribution is homogeneous (HCG) if and only if it has an interaction representation with no mixed quadratic interactions.

Inserting the interaction terms into (3.1) we get the following representation of the logarithm of the density:

$$\log f(i, y) = \sum_{d \subseteq \Delta} \lambda_d(i) + \sum_{d \subseteq \Delta} \sum_{\gamma \in \Gamma} \eta_d(i)_\gamma y_\gamma - \tfrac{1}{2} \sum_{d \subseteq \Delta} \sum_{\gamma, \mu \in \Gamma} \psi_d(i)_{\gamma\mu} y_\gamma y_\mu. \tag{3.2}$$

A CG-distribution is now said to be *a nearest-neighbour Gibbs* distribution w.r.t. G or *G-Gibbsian*, if it has an interaction representation with interaction terms satisfying

$$\begin{aligned} \lambda_d(i) &\equiv 0 \quad \text{unless } d \text{ is complete in } G, \\ \eta_d(i)_\gamma &\equiv 0 \quad \text{unless } d \cup \{\gamma\} \text{ is complete in } G, \\ \Psi_d(i)_{\gamma\mu} &\equiv 0 \quad \text{unless } d \cup \{\gamma, \mu\} \text{ is complete in } G. \end{aligned} \tag{3.3}$$

Thus a Gibbsian probability has an expansion with interaction terms only involving variables that are *neighbours*.

A key result in this section is the following version of the "Gibbs = Markov theorem," see, for example, Speed (1979) for a survey.

PROPOSITION 3.1. *A CG-distribution is $\mathcal{G}$-Markovian if and only if it is $\mathcal{G}$-Gibbsian.*

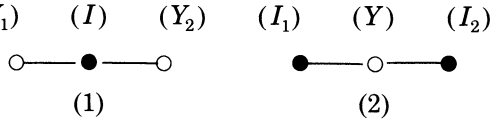
The general theorem implies that a distribution with positive density is Markov if and only if the density factorizes into a product of functions that only depend on variables that are mutual neighbours. We have to show that the factorization so obtained splits up into separate factorizations of the constant, linear and quadratic terms. This is done in Appendix B.

The corollary below, however, follows as in the standard case and its proof is therefore omitted.

COROLLARY 3.2. *The following statements are equivalent for a CG-distribution*:

(i) *The distribution is $\mathcal{G}$-Markovian.*
(ii) $a \perp\!\!\!\perp b \mid c$ *whenever* a, b, c *are disjoint and* c *separates* a *from* b.
(iii) $\forall\, \alpha \in V$: $\{\alpha\} \perp\!\!\!\perp V \setminus (\mathrm{cl}(\alpha)) \mid \mathrm{adj}(\alpha)$.

As a continuation of the examples from the previous section, consider the following graphs:



(1) (2)

The conditional independence restriction in the graph (1) is $Y_1 \perp\!\!\!\perp Y_2 \mid I$ and this is by Proposition 3.1 equivalent to the condition $k^{(12)}(i) \equiv 0$ in (2.6) because there is no edge between (Y_1) and (Y_2). In the second example, the restriction $I_1 \perp\!\!\!\perp I_2 \mid Y$

is by Proposition 3.1 equivalent to the existence of the expansion

$$\log f(i_1, i_2, y) = \lambda^{(1)}(i_1) + \lambda^{(2)}(i_2) + \big(\eta^{(1)}(i_1) + \eta^{(2)}(i_2)\big)y - \tfrac{1}{2}\big(\psi^{(1)}(i_1) + \psi^{(2)}(i_2)\big)y^2. \tag{3.4}$$

In the homogeneous case, that is, $\psi^{(1)}(i_1) + \psi^{(2)}(i_2) \equiv \psi$ the conditional distribution of Y given (I_1, I_2) will have expectation $\alpha(i_1) + \beta(i_2)$, where $\eta(i_1) = \lambda^{(1)}(i_1)/\psi$, $\beta(i_2) = \eta^{(2)}(i_2)/\psi$ and variance $\sigma^2 = \psi^{-1}$, that is, leading to the *additive model* for a two-way classification.

An important fact about CG-distributions that are Markovian is related to their behaviour under marginalization and conditioning. In fact we have the following proposition.

PROPOSITION 3.3. *If a CG-distribution is $\mathcal{G}$-Markovian, then for all $A \subset V$ the conditional distribution of X_A given $X_B = x_B^*$, where $B = V \setminus A$ is CG and G_A-Markovian.*

PROOF. That the CG-property is preserved is Proposition 2.3. That the Markov property is preserved is immediate from the definition of this property. □

In general, the Markov property is not preserved under marginalization. Consider the example 2 above, where the marginal distribution of (I_1, I_2) has point probabilities

$$\begin{aligned} p(i_1, i_2) &= \sqrt{2\pi\sigma^2}\exp\{\lambda^{(1)}(i_1) + \lambda^{(2)}(i_2) + \psi(\alpha(i_1) + \beta(i_2))/2\} \\ &= \exp\{a(i_1) + b(i_2) + \psi\alpha(i_1)\beta(i_2)\}. \end{aligned}$$

As we see, I_1 and I_2 are *not* independent, but p contains a *multiplicative interaction term* $\psi\alpha(i_1)\beta(i_2)$ in its logarithmic expansion. It is of interest to notice that this is exactly the model considered by Goodman (1979, 1981) and others for contingency tables with *ordered categories*. We see here that the 'ordering' of the categories naturally occurs from increasing values of the 'row effects' $\alpha(i_1)$ and 'column effects' $\beta(i_2)$.

A comprehensive discussion of the marginalization problem for CG-distributions has been given by Frydenberg (1988). We shall here only give the most basic result based on the notion of a *strongly simplicial subset*; see Appendix A.

PROPOSITION 3.4. *If X has a CG-distribution which is $\mathcal{G}$-Markovian and $B = V \setminus A$ is strongly simplicial, then X_A has a CG-distribution and is $\mathcal{G}_A$-Markovian.*

PROOF. Let us first look at the Markov property. Suppose $\alpha, \beta \in A$ are nonadjacent. Since B is simplicial, $\mathrm{bd}(B)$ is complete and at least one of them, say α, must be in $A \setminus \mathrm{cl}(B)$. Therefore all paths in G away from α must intersect A. Hence $A \setminus \{\alpha, \beta\}$ separates $\{\alpha\}$ from $\{\beta\}$ in $\mathcal{G}$ and $\{\alpha\} \perp\!\!\!\perp \{\beta\} \mid A \setminus \{\alpha, \beta\}$, whereby the Markov property follows from (ii) of Corollary 3.2.

To check the distributional property, consider first the case $B \subseteq \Gamma$ in which case Proposition 2.1 applies. If $B \subseteq \Delta$ its *strong* simpliciality ensures $\mathrm{bd}(B) \subseteq \Delta$,

whereby $B \perp\!\!\!\perp \Gamma \mid \Delta \setminus B$ and Proposition 2.2 applies. Finally, if B contains both discrete and continuous vertices, form the graph $\mathcal{G}^* = (V, E^*)$, where

$$(\alpha, \beta) \in E^* \quad \Leftrightarrow \quad \big[(\alpha, \beta) \in E \text{ or } \{\alpha, \beta\} \subseteq \mathrm{cl}(B)\big].$$

Then $B \cap \Gamma$ is strongly simplicial in $\mathcal{G}^*$ such that we can use the results above, first on $B \cap \Gamma$ and then on $B \cap \Delta$. □

In our example 1, the only nonsimplicial vertex is (I), corresponding to the distribution of (Y_1, Y_2) being neither Gaussian nor Markovian, since the boundary $\mathrm{bd}(\{(I)\})$ is equal to $\{(Y_1), (Y_2)\}$ and this is not discrete, nor complete. In the example 2, no nontrivial strongly simplicial subsets exist. The vertex (I_1) has boundary $\{(Y)\}$ and since this is complete, (I_1) is simplicial, but it does not satisfy the second requirement—that it be discrete because (I_1) is—and similarly with (I_2). The vertex (Y) is not even simplicial.

4. Graphical models of Markov type. In the present section we shall discuss statistical models for discrete and continuous variables based on the distributions considered in Section 3.

Suppose that we have N observations of vector-random variables $X^{(1)}, \ldots, X^{(N)}$, each of these having a set of qualitative (discrete) and a set of real-valued (continuous) components, that is,

$$X^{(\nu)} = \big(I^{(\nu)}, Y^{(\nu)}\big) = \big(I_\delta^{(\nu)}, \delta \in \Delta, Y_\gamma^{(\nu)}, \gamma \in \Gamma\big)$$

where $I_\delta^{(\nu)}$ take values in $\mathcal{I}_\delta$ and $Y_\gamma^{(\nu)}$ take values in the set of real numbers.

For each undirected marked graph $\mathcal{G} = (\Delta \cup \Gamma, E)$ the *graphical model corresponding to* $\mathcal{G}$ is defined by assuming that $X^{(1)}, \ldots, X^{(N)}$ are independent and identically distributed according to a distribution P, which is unknown apart from the fact that it is a *$\mathcal{G}$-Markovian CG-distribution* (or a $\mathcal{G}$-Markovian HCG-distribution).

Corollary 3.2 ensures that the models can be interpreted in terms of a distributional assumption and conditional independence statements, where the latter can be read directly off the graph. The likelihood function becomes

$$\begin{aligned} \log L &= \sum_{\nu=1}^{N} g(i^{(\nu)}) + \sum_{\nu=1}^{N} h(i^{(\nu)})^{\mathrm{T}} y^{(\nu)} - \tfrac{1}{2} \mathrm{tr}\left[\sum_{\nu=1}^{N} K(i^{(\nu)}) y^{(\nu)} y^{(\nu)\mathrm{T}}\right] \\ &= \sum_{i \in \mathcal{I}} \big[g(i)n(i) + h(i)^{\mathrm{T}} S(i) - \tfrac{1}{2} \mathrm{tr}(K(i)SP(i))\big], \end{aligned} \tag{4.1}$$

where we have let

$$n(i) = \sum_{\nu:\, i^{(\nu)} = i} 1 = \text{the number of observations with } I^{(\nu)} \text{ equal to } i,$$

$$S(i) = \sum_{\nu:\, i^{(\nu)} = i} y^{(\nu)} = \text{the sum of the corresponding } y\text{-vectors},$$

$$SP(i) = \sum_{\nu:\, i^{(\nu)} = i} y^{(\nu)} y^{(\nu)^{\mathrm{T}}} = \text{the matrix of sums of squares and products of corresponding } y\text{-vectors}.$$

For HCG-distributions, K does not depend on i so that the matrices of sums of squares and products can be pooled over i to give $SP = \sum_{i \in \mathcal{I}} SP(i)$.

From expression (4.1) we see that the set $(n(i), S(i), SP(i), i \in \mathcal{I})$ is a sufficient statistic and that we are in an exponential family; cf. Barndorff-Nielsen (1978). By standard results and by writing the joint density of X as the product of the marginal density of I and the conditional density of Y for given I, we obtain explicit estimates for g, h and K in the *unrestricted* case, that is, when $\mathcal{G}$ is the complete graph.

In the restricted case we define

$$\mathcal{C}_\Delta = \text{the set of cliques in } (\Delta, E_\Delta),$$

$$\mathcal{C}_\Delta(\gamma) = \text{the set of subsets } c \text{ of } \Delta \text{ such that } c \cup \{\gamma\} \text{ is a clique in } \left(\Delta \cup \{\gamma\}, E_{\Delta \cup \{\gamma\}}\right),$$

$$\mathcal{C}_\Delta(\gamma, \mu) = \text{the set of subsets } c \text{ of } \Delta \text{ such that } c \cup \{\gamma\} \cup \{\mu\} \text{ is a clique in } \left(\Delta \cup \{\gamma\} \cup \{\mu\}, E_{\Delta \cup \{\gamma\} \cup \{\mu\}}\right)$$

and note that $\mathcal{C}_\Delta(\gamma) = \mathcal{C}_\Delta(\gamma, \gamma)$.

Standard arguments imply that also in the restricted case we have an exponential family with canonical sufficient statistics for $i_c \in I_c$, where we have let $g(i_c) = \sum_{j:\, j_c = i_c} g(i)$ for any arbitrary function g:

$$n(i_c), \qquad c \in \mathcal{C}_\Delta,$$

$$S(i_c)_\gamma,\ SP(i_c)_{\gamma\gamma}, \qquad \gamma \in \Gamma,\ c \in \mathcal{C}_\Delta(\gamma),$$

$$SP(i_c)_{\gamma\mu}, \qquad \{\gamma, \mu\} \in E_\Gamma,\ c \in \mathcal{C}_\Delta(\gamma, \mu).$$

These statistics can be arranged in a hierarchy as follows:

1. A set of *marginal tables of counts* $(n(i_c), i_c \in \mathcal{I}_c)$ corresponding to the cliques of (Δ, E_Δ).
2. For each continuous variable $\gamma \in \Gamma$, a set of *marginal tables of sums and sums of squares* $(S(i_c)_\gamma, SP(i_c)_{\gamma\gamma})$ corresponding to the cliques of $(\Delta \cup \{\gamma\}, E_{\Delta \cup \{\gamma\}})$ of form c $\cup \{\gamma\}$.
3. For each pair of variables $\{\gamma, \mu\} \in E_\Gamma$ a set of *marginal tables of sums of products* $SP(i_c)_{\gamma\mu}$ corresponding to the cliques of $(\Delta \cup \{\gamma, \mu\}, E_{\Delta \cup \{\gamma, \mu\}})$ of form $c \cup \{\gamma, \mu\}$.

In the HCG-case there is only one table of sums of squares and products, some of the products not being needed, that is, for $(\{\gamma, \mu\} \notin E_\Gamma)$.

Illustrating this by the homogeneous case of example 2 in the previous section, we have

$$\mathcal{C}_\Delta = \{\{(i_1)\}, \{(i_2)\}\}, \qquad \mathcal{C}_\Delta((y)) = \mathcal{C}_\Delta$$

and the sufficient statistics are

$$\{n(i_1), n(i_2), S(i_1), S(i_2), SS\},$$

that is, the row and column marginal counts and sums as well as the total sum of squares.

By standard exponential family theory the maximum likelihood estimates are uniquely given by equating the value of the sufficient statistics to their expectations.

Frydenberg and Edwards (1988) have developed an algorithm for solving these equations by a modification of methods of iterative proportional scaling; cf., for example, Darroch and Ratcliff (1972) and Speed and Kiiveri (1986). The algorithm is implemented in the program MIM, documented in Edwards (1987). The program is developed to analyse the models described here as well as their generalizations to so-called hierarchical mixed interaction models, Edwards (1989).

5. The order Markov property, CG-distributions and CG-regressions. Contrasted with Section 3 we shall here study Markov-type properties relative to an *oriented* graph $\mathcal{G}$, where the orientation is induced by a complete ordering $<$, that is, where $\mathcal{G} = \mathcal{G}^{<}$.

A distribution on $\mathcal{X} = \mathcal{I} \times \mathcal{Y}$ with strictly positive density is said to be *order Markovian w.r.t.* $\mathcal{G}$ or *$\mathcal{G}$-order Markovian* if it satisfies

$$\textbf{OM}: \quad \{\alpha\} \perp\!\!\!\perp \Pi(\alpha) \setminus \mathrm{adj}(\alpha) \mid \mathrm{adj}(\alpha),$$

where $\Pi(\alpha) = \{\mu \in V \mid \mu < \alpha\}$. A slight modification of **OM** is the *local causal Markov property*, so-called by Kiiveri, Speed and Carlin (1984). The above is certainly equivalent, which follows from the main theorem in this reference. A distribution on $\mathcal{X} = \mathcal{I} \times \mathcal{Y}$ satisfies the order Markov property if and only if $\alpha \in V$,

$$\mathcal{L}\left(X_\alpha \mid X_{\Pi(\alpha)} = x_{\Pi(\alpha)}\right) = \mathcal{L}\left(X_\alpha \mid X_{\mathrm{adj}(\alpha)} = x_{\mathrm{adj}(\alpha)}\right). \tag{5.1}$$

In terms of interpretation one can think of X_α as a response, of $\Pi(\alpha)$ as the possible influencing variables for X_α and of $\mathrm{adj}(\alpha)$ as the variables directly influencing X_α.

A *recursive univariate CG-regression* is a distribution on $\mathcal{X}$ such that the conditional distributions (5.1) are univariate CG-regressions as described in Section 2. If these are all homogeneous we use the term *recursive univariate HCG-regression*.

In general such distributions are not CG. We have, however, the following proposition.

PROPOSITION 5.1. *If the ordering is strongly reducible any $\mathcal{G}$-order Markovian univariate recursive CG-regression is a $\mathcal{G}$-Markovian CG-distribution and vice versa.*

PROOF. That the Markov properties coincide is the corollary to the main theorem of Kiiveri, Speed and Carlin (1984). That the classes of distributions also coincide can be seen by an induction argument using Propositions 2.4, 3.1

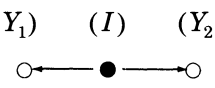
and 3.4. But the result, *as well as its converse* if $|\mathcal{J}_\delta| \geq 2$ is a special case of Proposition 8.2 below, see also Frydenberg (1988). □

REMARK 5.2. If we do not pay attention to the *Markovian* properties, it is not difficult to show *that an order Markovian recursive CG-regression is a CG-distribution if and only if* $\text{adj}(\delta) \subseteq \Delta$ *for all* $\delta \in \Delta$.

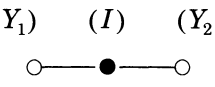
REMARK 5.3. Proposition 5.1 (together with its converse) contains as special cases results of Wermuth (1980) in the case $\Delta = \varnothing$ and Wermuth and Lauritzen (1983) in the case $\Gamma = \varnothing$. So do the results of Kiiveri, Speed and Carlin (1984) whereas they do not consider distributional properties outside the Gaussian case.

REMARK 5.4. Strongly reducible orderings exist by definition exactly for decomposable graphs so all Markovian CG-distributions on such graphs can be represented as order Markovian recursive CG-regressions after having chosen a suitable ordering.

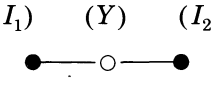
Continuing the examples, consider first the oriented graph



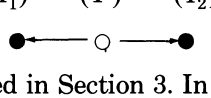
It displays a strongly reducible ordering and the corresponding order Markovian recursive regressions are the same as the Markovian CG-distributions for the corresponding undirected graph



and this graph is thus decomposable. On the other hand, the graph



does not admit a strongly reducible ordering and the distributions specified for the oriented version of example 2



are different from those specified in Section 3. In fact, in the homogeneous case, the directed version has log density which looks like

$$\begin{aligned} \log f(i_1, i_2, y) &= \log f(i_2 \mid y) + \log f(i_1 \mid y) + \log f(y) \\ &= \text{const.} + u^{(2)}(i_2) + v^{(2)}(i_2)y + u^{(1)}(i_1) + v^{(1)}(i_1)y \\ &\quad - \log \kappa^{(1)}(y) - \log \kappa^{(2)}(y) - \tfrac{1}{2}\psi y^2, \end{aligned} \tag{5.2}$$

which is different from that in (3.4), even if $\psi^{(1)}(i_1) + \psi^{(2)}(i_2) \equiv \psi$. The normalizing constants $\kappa^{(1)}$ and $\kappa^{(2)}$ are the 'troublemakers.'

6. Recursive graphical models. The models discussed in this section are based on the CG-regressions discussed in Section 2. We consider the same observational scheme as in Section 4, that is, N observations of vector random variables, each having a set of qualitative and a set of quantitative components. But now models are given by an *oriented* graph induced by a complete ordering $<$, that is, $\mathcal{G}^{<}$ such that the recursive *graphical model* corresponding to $\mathcal{G}$ is defined by assuming that the observations are realizations of $X^{(1)}, \ldots, X^{(N)}$, where $X^{(i)}$ are independent and identically distributed according to a distribution which is unknown, apart from the fact that it is a $\mathcal{G}$-order Markovian univariate recursive CG-regression (or HCG-regression).

Note that the recursive graphical models in the pure cases considered by Wermuth and Lauritzen (1983) and Wermuth (1980) are slightly more general than those defined here. The graphical chain models discussed in Section 8, however, are general enough to cover these as well.

The likelihood function factorizes into a product of conditional likelihood functions obtained by considering the conditional distribution of a variable X_α given its possible influences $X_{\Pi(\alpha)}$. Since by construction the parameters in each of these conditional distributions vary freely and independently of those in other conditional distributions, the likelihood function can be maximized by maximizing each factor separately. Let us derive an expression for such a conditional likelihood function in the case of a discrete variable $\alpha = \delta$ using expressions given in Section 2:

$$\begin{aligned}\log L &= \sum_{\nu=1}^{N} \log P\Big\{ X_\delta^{(\nu)} = i_\delta^{(\nu)} \mid X_{\mathrm{adj}(\delta)}^{(\nu)} = \Big(i_{\Delta\,\mathrm{bd}(\delta)}^{(\nu)},\, y_{\Gamma\,\mathrm{bd}(\delta)}^{(\nu)} \Big) \Big\} \\ &= \sum_{i_d \in \mathcal{I}_{\delta \cup \Delta\,\mathrm{bd}(\delta)}} \big[u^\delta(i_\delta \mid j) n(i_d) + v^\delta(i_\delta \mid j) S_\delta(i_d) + \mathrm{tr}\big(W^\delta(i_\delta \mid j) SP_\delta(i_d) \big) \big] \\ &\quad - \sum_{\nu=1}^{N} \log \kappa^\delta\Big(i_{\Delta\,\mathrm{bd}(\delta)}^{(\nu)},\, y_{\Gamma\,\mathrm{bd}(\delta)}^{(\nu)} \Big),\end{aligned}$$

where we have used the notation $(i_\delta, j) = i_d$ and

$n(i_d)$ = the number of observations with $i_{\delta \cup \Delta\,\mathrm{bd}(\delta)}^{(\nu)} = i_d$,

$S_\delta(i_d)$ = the sum of the values of $y_\gamma^{(\nu)}$, $\gamma \in \Gamma\,\mathrm{bd}(\delta)$, for those ν, where $i_{\delta \cup \Delta\,\mathrm{bd}(\delta)}^{(\nu)} = i_d$,

$SP_\delta(i_d)$ = the matrix of sums of squares of products of the same $y_\gamma^{(\nu)}$-values.

As seen from the above expression (which is also well known) we have an exponential family likelihood with (conditionally) sufficient statistics

$$\big[n(i_d), S_\delta(i_d), SP_\delta(i_d), i_d \in \mathcal{I}_{\delta \cup \Delta\,\mathrm{bd}(\delta)} \big].$$

In the general case [$\Gamma\,\mathrm{bd}(\delta) \neq \varnothing$] there is no reduction in the term involving the normalizing constant κ^δ and we have no explicit formula for the maximum likelihood estimates. Also there is in general no *jointly* sufficient reduction, because the terms involving κ^δ will not have the status of normalization

constants in the joint likelihood. Then the full data set might be needed to calculate maximum likelihood estimates.

In the case $\Gamma \operatorname{bd}(\delta = \varnothing)$, however, the likelihood function reduces to

$$\log L = \sum_{i_d \in \mathscr{I}_{\delta \cup \Delta \operatorname{bd}(\delta)}} u^{\delta}(i_\delta \mid j) n(i_d) - \sum_{j \in \mathscr{I}_{\Delta \operatorname{bd}(\delta)}} \log \kappa^{\delta}(j) n(j)$$

and this is maximized by letting

$$\hat{u}^{\delta}(i_\delta \mid j) = \log \frac{n(i_d)}{n(j)} \quad \text{and} \quad \hat{\kappa}^{\delta}(j) = 1.$$

In the homogeneous case, we get essentially the same phenomena as above, just that now the $SP_\delta(i_d)$-terms are not needed in the set of sufficient statistics.

The regression problems for $\gamma \in \Gamma$ are univariate linear model problems with well-known explicit solutions.

To summarize, the joint likelihood function obtained by multiplying conditional likelihood functions together, has in general no nice properties of sufficiency and exponential family type. But each of the conditional likelihood functions has and a unique maximum likelihood estimate can be obtained by maximizing each factor. The maximization problems have an explicit solution in the continuous cases but we know only explicit solutions to the discrete cases when $\Gamma \operatorname{bd}(\delta) = \varnothing$. As a consequence, *the pure cases* ($\Gamma = \varnothing$ *or* $\Delta = \varnothing$) *always have an explicit solution.*

Note especially that when the ordering is strongly reducible, $\Gamma \operatorname{bd}(\delta) = \varnothing$ for all $\delta \in \Delta$ and therefore the maximum likelihood estimates of the parameters can be found explicitly, see the next section.

7. Decomposable graphical models. Consider a graphical model of Markov type given by a decomposable undirected marked graph $\mathscr{G} = (\Delta \cup \Gamma, E)$. Since it is of the type considered in Section 4, we have an exponential family structure for the joint likelihood and sufficient reductions to marginal tables of counts, sums and sums of squares and products as described in that section.

By definition, a graph is decomposable if and only if a strongly reducible ordering $<$ exists. By Proposition 5.1, any distribution in the family considered is also an order Markovian CG-regression and vice versa, whereby we conclude that the model is equivalent to the *recursive* graphical model given by the oriented graph $\mathscr{G}^{<}$. Since the ordering is strongly reducible, $\Gamma \operatorname{bd}(\delta) = \varnothing$ for all $\delta \in \Delta$, and we can obtain *explicit* estimates of the parameters in the recursive graphical model. Using the equivalence between the models once more, we can obtain *explicit estimates* of the parameters in the graphical model given by $\mathscr{G}$. It thus follows that such models have nice properties in terms of sufficient reductions as well as explicit solutions to the estimation problem. In general, a decomposable graph will admit several strongly reducible orderings and it thus follows that although their interpretations are different, *all the corresponding recursive graphical models are identical* and equal to the graphical model given by the undirected graph $\mathscr{G}$.

Each strongly reducible ordering of a decomposable graph represents a *recursive dependence structure*, thus characterizing decomposable graphical models as the subclass of graphical Markovian models for which an interpretation with recursively ordered responses applies.

We mention that it has been shown, Lauritzen (1985) that the likelihood ratio for testing one decomposable model versus another can be partitioned as a product of likelihood ratios for well-known linear models and/or conditional independence tests in contingency tables. This has been used by Williams (1976) in the discrete case and by Porteous (1985a) in the continuous case to obtain Bartlett corrections.

8. Graphical chain models. The notion of a Markovian graphical model and a recursive graphical model can be unified in the notion of a *graphical chain model* to be briefly described below. While a Markovian graphical model contains no arrows (undirected graph) and a recursive graphical model contains solely arrows (oriented graph) in its picture, the picture of a graphical chain model will in general contain both.

We consider a chain graph $\mathcal{G} = \mathcal{G}^{<}$ with chain $V(1), \ldots, V(T)$ that we here shall refer to as a *dependence chain*.

We now let $D(\alpha) = \{\beta \mid \beta < \alpha \text{ and } \beta \neq \alpha\}$ and consider the *chain Markov property*

$$\mathbf{CM}: \quad \alpha \perp\!\!\!\perp D(\alpha) \setminus \operatorname{adj}(\alpha) \mid \operatorname{adj}(\alpha).$$

This specializes to the usual Markov property (**M**) if $T = 1(\mathcal{G} = \mathring{\mathcal{G}})$ and the order Markov property **OM** if $|V(t)| \equiv 1$.

A corresponding class of distributions is the class of *recursive multivariate CG-regressions or HCG-regressions*, that is, where the conditional distribution of $X_{V(t)}$ given $X_{W(t)}$, where

$$W(t) = \bigcup_{l<t} V(l),$$

is of the type considered in Section 3.

The graphical chain model given by $\mathcal{G}$ and the dependence chain $V(1), \ldots, V(T)$ is now obtained by assuming the observations to be realizations of independent identically distributed random variables, with a distribution being unknown, apart from the fact that it is a recursive multivariate CG-regression (or HCG-regression) satisfying the chain Markov property **CM**. Frydenberg (1986) has shown that two chain models with the same graph but different dependence chains are identical so that the model in fact is determined by the graph $\mathcal{G}$ alone and reference to the dependence chain can be omitted. As in the previous section, the likelihood function is most conveniently analysed by considering each of the conditional likelihood functions obtained from the conditional distributions of $x_{V(t)}$ given $X_{W(t)}$. We shall abstain from giving the details.

We have seen (Proposition 5.1) that under certain circumstances the models for symmetric associations and the directional models coincide. That is of independent interest and leads, for example, to the identification of the class of

decomposable models as in the previous section. But results on such equivalences can also be useful for a variety of other purposes, ranging from computational shortcuts in the fitting of models to aspects of resolving controversies about the interpretation of data, see, for example, Wermuth and Lauritzen (1983, 1989) for a discussion. A nonstandard example of an application is the recent work of Lauritzen and Spiegelhalter (1988) where ideas along these lines have been used to develop methods for efficient calculations with probabilities in expert systems. Here we just briefly state and illustrate the main results.

PROPOSITION 8.1. *If* $|\mathscr{I}_\delta| \geq 2$ *for all* δ, *the graphical chain model given by* $\mathscr{G}$ *is equivalent to the Markovian graphical model given by* $\mathring{\mathscr{G}}$, *if and only if the dependence chain is strongly reducible.*

The results, in this generality, is due to Frydenberg (1988) and stated and proved there as Proposition 5.6. We therefore omit the proof. As an illustration of the use of the result, the models below are equivalent:

(the dependence chain is illustrated by boxes) implying, for example, that estimation in the model to the left can be performed in the model to the right, that is, ignoring the response structure, whereas this is not the case for the models

The model to the left specifies, for example, *marginal* independence of the variables in the left box, whereas the model to the right specifies *conditional* independence of the two, given the remaining variables. See Wermuth and Lauritzen (1989) for a wide range of similar examples.

The corresponding conditions for coincidence between recursive models and graphical chain models are briefly stated below, in the case where the orderings involved in the recursive model ($\prec$) and in the chain model ($<$) are assumed *compatible*, that is, no arrows in one graph are reversed in the other.

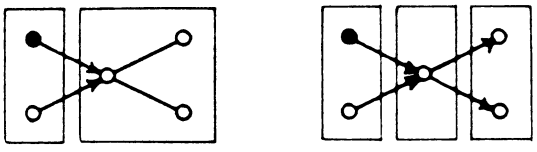
PROPOSITION 8.2. *If a complete ordering $\prec$ is locally strongly reducible and compatible, the graphical chain model given by $\mathcal{G}$ and the recursive graphical model given by $\mathcal{G}^{\prec}$ are equivalent.*

PROPOSITION 8.3. *If a complete ordering $\prec$ is strongly reducible and compatible, the graphical chain model given by $\mathcal{G}$, the Markovian graphical model given by $\mathring{\mathcal{G}}$ and the recursive graphical model given by $\mathcal{G}^{\prec}$ are all equivalent.*

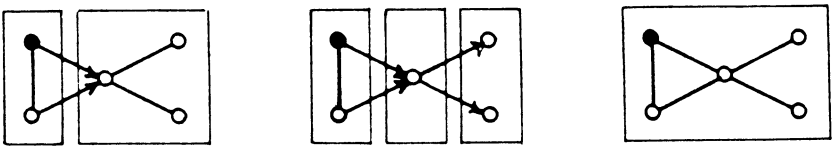
PROOF. Propositions 8.2 and 8.3 are proved by induction on T and repeated use of Proposition 5.1. □

We believe the conditions above to be necessary as well although no formal proof has been established.

A short illustration of this: The two models below are equivalent



by Proposition 8.2 and the three models below are by Proposition 8.3,



implying that the graph to the far right is decomposable.

APPENDIX A

Graph theory. Graph-theoretic aspects of the models considered in the present article were first discussed by Lauritzen and Wermuth (1984). Since then, a thorough study of this has been performed by Leimer (1985, 1989). This had led to considerable improvements of the terminology and of the general understanding of such graphs. We here just give a minimal treatment and refer the reader to the above for details.

Our graphs are *simple*, that is, there are no loops or multiple edges. The vertices are *marked*, reflecting the necessity to keep track of two kinds of variables.

A *marked graph* consists of a finite set V of *vertices*, partitioned into two disjoint subsets $V = \Delta \cup \Gamma$, and a collection E of *edges* being a subset of the set of ordered pairs of distinct elements of V. We write $\mathcal{G} = (V, E)$ or $\mathcal{G} = (\Delta \cup \Gamma, E)$.

It is important that the properties of our graphs, which refer specifically to the two types of vertices are not symmetric in Δ and Γ. Vertices in Δ are supposed to represent *discrete* variables and Γ *continuous* variables, and they do play different roles. We shall use *discrete* and *continuous* for the vertices Δ and Γ, respectively. If either Δ or Γ is empty, the graph is *pure*.

We represent the graph by a picture with *d*iscrete vertices as *d*ots and *c*ontinuous vertices as *c*ircles.

An edge $(\alpha, \beta) \in E$ is represented by an *arrow* from α to β if $(\beta, \alpha) \notin E$, and by a line between α and β if both $(\alpha, \beta) \in E$ and $(\beta, \alpha) \in E$. Examples are in the main body of the article.

A graph is called *undirected* if there are no arrows in the picture, otherwise we call it a *directed* graph.

A graph is called *oriented* if the picture has solely arrows and no lines.

The *symmetrization* of $\mathring{\mathcal{G}}$ of a graph $\mathcal{G}$ is obtained from $\mathcal{G}$ by substituting lines for arrows all over.

A special type of directed graph occurs when the vertex set is partitioned into an ordered sequence of subsets $V(1), \ldots, V(T)$ to be called a *chain*.

The chain induces a partial order $<$ on the vertices as

$$\alpha < \beta \quad \Leftrightarrow \quad \exists\, s, t, \qquad \alpha \in V(s), \qquad \beta \in V(t), \qquad s \leq t.$$

We then define for $\mathcal{G} = (V, E)$ the induced *chain graph* $\mathcal{G}^{<} = (V, E^{<})$ as

$$(\alpha, \beta) \in E^{<} \quad \Leftrightarrow \quad (\alpha, \beta) \in E \quad \wedge \quad \alpha < \beta.$$

The chain graph has lines between vertices in the same chain element and arrows between vertices in different elements, all arrows pointing from low to high.

If $|\, V(t)\,| \equiv 1$, we have a complete ordering of the vertices and $\mathcal{G}^{<}$ will be an oriented graph. If only $V(1)$ has more than one element, $\mathcal{G}^{<}$ will be what Kiiveri, Speed and Carlin (1984) call a *recursive causal graph*.

The *subgraph induced by a subset* $A \subseteq V$ of the vertex set given as $\mathcal{G}_A = (A, E_A)$ where $E_A = E \cap (A \times A)$.

A graph is *complete* if all pairs of distinct vertices are connected with an arrow or a line.

A subset is *complete* if it induces a complete subgraph. A maximal (w.r.t. inclusion) complete subset is called a *clique*.

To a graph $\mathcal{G}$ corresponds its *adjacency function*, given as

$$\alpha \in \mathrm{adj}(\beta) \quad \Leftrightarrow \quad (\alpha, \beta) \in E,$$

that is, the vertices α *adjacent* to β are those being the starting point of arrows pointing toward β or lines between α and β. Note that this is reversed compared to Golumbic (1980).

If $\mathcal{G}$ is undirected we have

$$\alpha \in \text{adj}(\beta) \quad \Leftrightarrow \quad \beta \in \text{adj}(\alpha),$$

and α and β are called *adjacent* or *neighbours*.

For $A \subseteq V$ we define its *boundary* and *closure* as

$$\text{bd}(A) = \bigcup_{\alpha \in A} \text{adj}(\alpha) \cap (V \setminus A), \qquad \text{cl}(A) = A \cup \text{bd}(A),$$

the boundary of A thus being all vertices adjacent to some vertex $\alpha \in A$.

With special reference to marked graphs we also define the *discrete and continuous boundaries* as

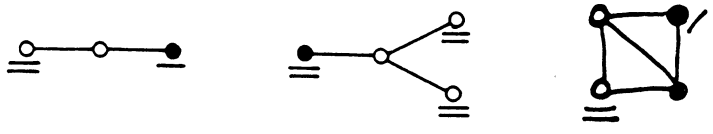
$$\Delta\,\text{bd}(A) = \text{bd}(A) \cap \Delta, \qquad \Gamma\,\text{bd}(A) = \text{bd}(A) \cap \Gamma.$$

A *path* of *length n* from α to β is a sequence of vertices $\alpha = \alpha_0, \alpha_1, \ldots, \alpha_n = \beta$, such that $(\alpha_{i-1}, \alpha_i) \in E$ for all $i = 1, \ldots, n$, and all vertices except possibly α and β are distinct. If $\alpha = \beta$ the path is a *cycle*.

A cycle is *chordless* if only consecutive elements are joined with edges.

Two disjoint subsets A and B are said to be *separated* by a subset C (disjoint from A and B) if all paths from A to B contain vertices from C.

A vertex α is called *simplicial* if its adjacency set is complete and with special reference to marked graphs, a vertex α is *strongly simplicial* if it is simplicial and $\alpha \in \Gamma$ or $\text{adj}(\alpha) \subseteq \Delta$. In words, a discrete simplicial vertex is supposed to have only discrete vertices in its adjacency set. Note the asymmetry in this definition, and that, *if $\mathcal{G}$ is pure, any simplicial vertex is strongly simplicial.* The vertices underlined are simplicial in the graphs below and those double underlined are strongly simplicial:



We generalize this notion to subsets A by saying that A is *simplicial* if its boundary $\text{bd}(A)$ is complete and *strongly simplicial* if further $A \subseteq \Gamma$ or $\text{bd}(A) \subseteq \Delta$.

Following Frydenberg (1986, 1988) A is called a (*strongly*) *simplicial collection* if all connected components $A_1, \ldots, A_p$ of the subgraph $\mathcal{G}_A$ are (strongly) simplicial in $\mathcal{G}$.

An ordering induced by a chain $V(1), \ldots, V(T)$ is called (*strongly*) *reducible* if all chain elements are (strongly) simplicial collections in the induced chain graph $\mathcal{G}^{<}$. This extends the notion of a reducible numbering as used by Wermuth (1980) and Wermuth and Lauritzen (1983).

A complete ordering $\prec$ of the vertices in a chain graph is said to be *compatible* if

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

where the right inequality sign refers to the ordering given by the chain graph.

A compatible ordering is *locally reducible* if $\rho \in V(t)$ and $\alpha, \beta \in \text{adj}^{\prec}(\rho)$ implies

$$\alpha \in \text{adj}^{\prec}(\beta) \quad \text{or} \quad \beta \in \text{adj}^{\prec}(\alpha) \quad \text{or} \quad \alpha, \beta \notin V(t).$$

It is *locally strongly reducible* if also

$$\rho \in \Gamma \quad \text{or} \quad \text{adj}^{\prec}(\rho) \cap V(t) = \varnothing.$$

Of special interest to us is the class of *triangulated graphs*, these being undirected graphs with no chordless cycles of length ≥ 4. These have been extensively studied by Dirac (1961) and many authors, occasionally under other names. For further information on this issue see Berge (1973), Golumbic (1980), Lauritzen, Speed and Vijayan (1984) and Darroch, Lauritzen and Speed (1980) together with references therein. The following results can be found in Golumbic (1980).

PROPOSITION A.1. *An undirected graph is triangulated if and only if there exists a reducible ordering of the vertices.*

PROPOSITION A.2. *Any triangulated graph has at least one simplicial vertex.*

PROPOSITION A.3. *If $\mathcal{G}$ is triangulated and A is a subset of V, then $\mathcal{G}_A = (A, E_A)$ is triangulated.*

These results automatically give an algorithm for recognizing triangulated graphs. First, look for a simplicial vertex. If such a vertex does not exist, the graph is not triangulated. Otherwise, remove the simplicial vertex α by forming the graph $\mathcal{G}_{V \setminus \{\alpha\}}$. Repeat the procedure on the subgraph of $\mathcal{G}$. Either we get at some stage a graph without a simplicial vertex and the graph is not triangulated, or we end up with the empty graph. Introduce now the ordering of V,

$$\alpha < \beta \quad \Leftrightarrow \quad \beta \text{ was removed before } \alpha.$$

This ordering will necessarily be reducible since the adjacency set of any vertex in $\mathcal{G}^{<}$ exactly will be its adjacency set in the subgraph that is left over just before it has been removed. This algorithm is inefficient in terms of computing time but a fast algorithm exists; see Tarjan and Yannakakis (1984). Motivated by these algorithms, we state

DEFINITION A.4. *An undirected marked graph $\mathcal{G} = (\Delta \cup \Gamma, E)$ is called decomposable if there exists a strongly reducible ordering of $\Delta \cup \Gamma$.*

The definition fits well with an algorithm of the type just described but also the fast algorithms can be generalized; see Leimer (1989).

In the pure cases an undirected graph is decomposable if and only if it is triangulated; cf. Proposition A.1.

APPENDIX B

CG-Markov = CG-nearest-neighbour Gibbs. The proof of Proposition 3.1 is basically a modification of the standard proof of the result in the discrete case. We first need a lemma.

LEMMA B.1 (Möbius inversion). *Let H and K be functions defined on the subsets of a finite set A and taking values in an Abelian group. Then the following are equivalent*:

$$\text{(i)} \qquad \forall\, a \subseteq A: \qquad H(a) = \sum_{b \subseteq a} J(b),$$

$$\text{(ii)} \qquad \forall\, a \subseteq A: \qquad J(a) = \sum_{b \subseteq a} (-1)^{|a \setminus b|} H(b).$$

For a proof, see, for example, Aigner (1979).

Let f be the density of a CG-distribution and let its logarithm be expressed as in (3.2) with interaction terms satisfying (3.3), that is, with only interactions among neighbours. Let $\alpha \notin \text{adj}(\beta)$. We have to show

$$\alpha \perp\!\!\!\perp \beta \mid V \setminus \{\alpha, \beta\}.$$

But this follows from (3.2) and from conditional independence, since no interaction terms involving *both* x_α and x_β will be present in (3.2) thus giving a factorization of the density f into a function not depending on x_α and one not depending on x_β.

The reverse implication demands a somewhat more complicated argument, since we have to *construct* a representation (3.2) and afterwards show that it satisfies (3.3). This is where the Möbius inversion lemma shall prove useful.

First we define elements of $\mathcal{I}$ depending on $d \subseteq \Delta$ as follows: Choose a *fixed* but arbitrary $i^* \in \mathcal{I}$. For $i \in \mathcal{I}$, let $i(d) \in \mathcal{I}$ be given as the "substituted" element

$$i(d)_\delta = \begin{cases} i_\delta & \text{if } \delta \in d, \\ i^*_\delta & \text{if } \delta \notin d. \end{cases}$$

Thus, for example, we have $i(\Delta) = i$, $i(\varnothing) = i^*$.

Let a density f of a CG-distribution be given by the expression (3.1) and define the functions

$$\text{(B.1)} \qquad \rho_d(i) = g(i(d)), \qquad \xi_d(i) = h(i(d)), \qquad \Phi_d(i) = K(i(d))$$

and further for $a \subseteq \Delta$,

$$\begin{aligned}\lambda_a(i) &= \sum_{d \subseteq a} (-1)^{|a \setminus d|} \rho_d(i),\\ \eta_a(i) &= \sum_{d \subseteq a} (-1)^{|a \setminus d|} \xi_d(i),\\ \Psi_a(i) &= \sum_{d \subseteq a} (-1)^{|a \setminus d|} \Phi_d(i).\end{aligned} \tag{B.2}$$

For fixed $i \in I$, the functions entering into (B.2) can be considered as functions on the subsets of Δ into the groups $(\mathbb{R}, +)$, $(\mathbb{R}^k, +)$, $(\mathbb{R}^{k \times k}, +)$, and Lemma B.1 applies. In the special case where $d = \Delta$, $i(\Delta) = i$, we get from (B.1)

$$\begin{aligned}\rho_\Delta(i) = g(i) &= \sum_{d \subseteq \Delta} \lambda_d(i),\\ \xi_\Delta(i) = h(i) &= \sum_{d \subseteq \Delta} \eta_d(i),\\ \Phi_\Delta(i) = K(i) &= \sum_{d \subseteq \Delta} \Psi_d(i),\end{aligned}$$

and we have constructed a representation of the form (3.2) for the density f. What remains to be shown is that also (3.3) is satisfied. The necessary trick is to identify terms in the expansions with particular values of the density and then to use (B.2) and the Markov property to see that zeros occur.

We first define the vector $e(\alpha) \in \mathbb{R}^\Gamma$ as that with one in position α and zero elsewhere.

We then have the following expressions for the terms in (B.1):

$$\rho_d(i) = \log f(i(d), 0), \tag{B.3}$$

$$\begin{aligned}\xi_d(i)_\alpha = {} & 2 \log f(i(d), e(\alpha)) - \tfrac{1}{2} \log f(i(d), 2e(\alpha))\\ & - \tfrac{3}{2} \log f(i(d), 0)\end{aligned} \tag{B.4}$$

and, if $\alpha \neq \beta$,

$$\begin{aligned}\Phi_d(i)_{\alpha\beta} = {} & \log f(i(d), e(\alpha)) - \log f(i(d), e(\alpha) + e(\beta))\\ & + \log f(i(d), e(\beta)) - \log f(i(d), 0),\end{aligned} \tag{B.5}$$

whereas

$$\begin{aligned}\Phi_d(i)_{\alpha\alpha} = {} & 2 \log f(i(d), e(\alpha)) - \log f(i(d), 2e(\alpha))\\ & - \log f(i(d), 0).\end{aligned} \tag{B.6}$$

Suppose now that d is not complete. Then there exist $\delta, \varepsilon \in d$ nonadjacent, which by the Markov property implies that $\{\delta\} \perp\!\!\!\perp \{\varepsilon\} \mid V \setminus \{\delta, \varepsilon\}$. Using now

(B.2) and (B.3), we get

$$\begin{aligned}\lambda_d(i) &= \sum_{a\subseteq d} (-1)^{|d\setminus a|}\rho_a(i)\\ &= \sum_{a\subseteq d\setminus\{\delta,\varepsilon\}} (-1)^{|d\setminus a|}\left[\rho_{a\cup\{\delta,\varepsilon\}}(i) - \rho_{a\cup\{\varepsilon\}}(i) - \rho_{a\cup\{\delta\}}(i) + \rho_a(i)\right]\\ &= \sum_{a\subseteq d\setminus\{\delta,\varepsilon\}} (-1)^{|d\setminus a|}\log\frac{f(i(a\cup\{\delta,\varepsilon\}),0)f(i(a),0)}{f(i(a\cup\{\delta\}),0)f(i(a\cup\{\varepsilon\}),0)}.\end{aligned}$$

In each of the terms in the above ratios, we have $X_a = i_a$, $X_b = i_b^*$, $X_\Gamma = 0$, where $b = \Delta \setminus (a \cup \{\delta, \varepsilon\})$. Conditioning on this, we obtain that the ratios are equal to

$$\begin{aligned}&\Big(P\{X_{\{\delta,\varepsilon\}} = (i_\delta, i_\varepsilon) | X_a = i_a,\ X_b = i_b^*,\ X_\Gamma = 0\}\\ &\quad\times P\{X_{\{\delta,\varepsilon\}} = (i_\delta^*, i_\varepsilon^*) | X_a = i_a,\ X_b = i_b^*,\ X_\Gamma = 0\}\Big)\\ &\quad\div\Big(P\{X_{\{\delta,\varepsilon\}} = (i_\delta, i_\varepsilon^*) | X_a = i_a,\ X_b = i_b^*,\ X_\Gamma = 0\}\\ &\quad\times P\{X_{\{\delta,\varepsilon\}} = (i_\delta^*, i_\varepsilon) | X_a = i_a,\ X_b = i_b^*,\ X_\Gamma = 0\}\Big)\end{aligned}$$

This ratio is equal to one by conditional independence, and thus $\lambda_d(i) \equiv 0$.

Using the same kind of argument for η_d and Ψ_d gives $\eta_d \equiv \Psi_d \equiv 0$ just that the corresponding term inside square brackets has to be split into three or four terms depending on whether (B.4), (B.5) or (B.6) is used.

If d is complete but $d \cup \{\gamma\}$ is not, there must be a $\delta \in d$ with $\delta \notin \text{adj}(\gamma)$. Then

$$\eta_d(i)_\gamma = \sum_{a\subseteq d\setminus\{\delta\}} (-1)^{|d\setminus a|}\big(\xi_a(i)_\gamma - \xi_{a\cup\{\delta\}}(i)_\gamma\big).$$

Using now (B.4) and conditional independence, we obtain $\eta_d(i)_\gamma \equiv 0$, and similarly for ψ_d.

Finally, if $d \cup \{\gamma\}$ is complete and also $d \cup \{\mu\}$ but not $d \cup \{\gamma, \mu\}$, we must have $\gamma \notin \text{adj}(\mu)$ and thus $\{\gamma\} \perp\!\!\!\perp \{\mu\} \mid V \setminus \{\gamma, \mu\}$. From (B.5) we get $\Phi_d(i)_{\gamma\mu} \equiv 0$ and by (B.2)

$$\Psi_d(i)_{\gamma\mu} = \sum_{a\subseteq d} (-1)^{|d\setminus a|}\Phi_a(i)_{\gamma\mu} \equiv 0.$$

Proposition 3.1 has been proved. □

Acknowledgments. We are indebted to Morten Frydenberg, Finn Søholm Larsen, Hanns-Georg Leimer, Klaus Rostgaard and several referees for critical readings of earlier versions of this article leading to the correction of errors and other improvements.

REFERENCES

AIGNER, M. (1979). *Combinatorial Theory*. Springer, Berlin.

ASMUSSEN, S. and EDWARDS, D. (1983). Collapsibility and response variables in contingency tables. *Biometrika* **70** 567–578.

BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory.* Wiley, New York.

BERGE, C. (1973). *Graphs and Hypergraphs.* North-Holland, Amsterdam.

BLALOCK, H. M., JR. (1971). *Causal Models in the Social Sciences.* MacMillan, London.

DARROCH, J. N. and RATCLIFF, D. (1972). Generalized iterative scaling for log-linear models. *Ann. Math. Statist.* **43** 1470–1480.

DARROCH, J. N. and SPEED, T. P. (1983). Additive and multiplicative models and interactions. *Ann. Statist.* **11** 724–738.

DARROCH, J. N., LAURITZEN, S. L. and SPEED, T. P. (1980). Markov fields and log-linear interaction models for contingency tables. *Ann. Statist.* **8** 522–539.

DAWID, A. P. (1979). Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. Ser. B* **41** 1–31.

DAWID, A. P. (1980). Conditional independence for statistical operations. *Ann. Statist.* **8** 598–617.

DEMPSTER, A. P. (1972). Covariance selection. *Biometrics* **28** 157–175.

DEMPSTER, A. P. (1973). Aspects of the multinomial logit model. *Multivariate Analysis III* (P. R. Krishnaiah, ed.) 129–142. Academic, New York.

DIRAC, G. A. (1961). On rigid circuit graphs. *Abh. Math. Sem. Univ. Hamburg* **25** 71–76.

EDWARDS, D. (1987). A guide to MIM. Research Report 87/1, Statist. Res. Unit, Univ. Copenhagen.

EDWARDS, D. (1988). Graphical modelling in multivariate analysis. *Proc. First Internat. Conf. on Statistical Computing, Izmir, Turkey.* To appear.

EDWARDS, D. (1989). Hierarchical interaction models. *J. Roy. Statist. Soc. Ser. B* **51**. To appear.

FRYDENBERG, M. (1986). Blandede interaktionsmodeller, kausale modeller, kollapsibilitet og estimation. Thesis, Aarhus Univ., Statistiske Interna 42.

FRYDENBERG, M. (1988). Marginalization and collapsibility in graphical association models. Research Report 166, Dept. Theoretical Statistics, Aarhus Univ.

FRYDENBERG, M. and EDWARDS, D. (1988). A modified iterative proportional scaling algorithm for estimation in regular exponential families. Research Report 167, Dept. Theoretical Statistics, Aarhus Univ.

GOLDBERGER, A. S. and DUNCAN, O. D., eds. (1973). *Structural Equation Models in the Social Sciences.* Seminar, New York.

GOLUMBIC, M. C. (1980). *Algorithmic Graph Theory and Perfect Graphs.* Academic, London.

GOODMAN, L. A. (1973). The analysis of multidimensional contingency tables when some variables are posterior to others. A modified path analysis approach. *Biometrika* **60** 179–192.

GOODMAN, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* **74** 537–552.

GOODMAN, L. A. (1981). Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* **76** 320–334.

HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–960.

JÖRESKOG, K. G. (1977). Structural equation models in the social sciences: Specification, estimation and testing. In *Applications of Statistics* (P. Krishnaiah, ed.) 265–287. North-Holland, Amsterdam.

KIIVERI, H. T. (1983). A unified theory of causal models. Ph.D. dissertation, Univ. Western Australia.

KIIVERI, H. and SPEED, T. P. (1982). Structural analysis of multivariate data: A review. In *Sociological Methodology* (S. Leinhart, ed.) 209–289. Jossey-Bass, San Francisco.

KIIVERI, H., SPEED, T. P. and CARLIN, J. B. (1984). Recursive causal models. *J. Austral. Math. Soc. A* **36** 30–52.

KRZANOWSKI, W. J. (1983). Distance between populations using mixed continuous and categorical variables. *Biometrika* **70** 235–243.

LAURITZEN, S. L. (1985). Test of hypotheses in decomposable mixed interaction models. Research Report R-85-11, Inst. Electronic Systems, Aalborg Univ.

LAURITZEN, S. L. and SPIEGELHALTER, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **50** 157–224.

LAURITZEN, S. L. and WERMUTH, N. (1984). Mixed interaction models. Research Report R-84-8, Inst. Electronic Systems, Aalborg Univ.

LAURITZEN, S. L., SPEED, T. P. and VIJAYAN, K. (1984). Decomposable graphs and hypergraphs. *J. Austral. Math. Soc. A* **36** 12–19.

LEIMER, H.-G. (1985). Strongly decomposable graphs and hypergraphs. Thesis, Ber. z. Stochastik u. verw. Gebiete 85-1, Univ. Mainz.

LEIMER, H.-G. (1989). Triangulated graphs with marked vertices. In *Graph Theory in Memory of G. A. Dirac* (L. D. Andersen et al., eds.). *Ann. Discrete Math.* **41** 311–324. North-Holland, Amsterdam.

LITTLE, R. J. A. and SCHLUCHTER, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* **72** 497–512.

OLKIN, I. and TATE, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Ann. Math. Statist.* **32** 448–465.

PORTEOUS, B. T. (1985a). Improved likelihood ratio statistics for covariance selection models. *Biometrika* **72** 473–475.

PORTEOUS, B. T. (1985b). Properties of log linear and covariance selection models. Ph.D. dissertation, Cambridge Univ.

SCHUMANN, R. (1986). Arbeitsgedächtnis—ein entwicklungspsychologisches Konzept untersucht am Beispiel multipler Bildvergleiche. Dissertation, Inst. Psychology, Univ. Mainz.

SIMON, H. A. (1957). *Models of Man.* Wiley, New York.

SPEED, T. P. (1979). A note on nearest-neighbour Gibbs and Markov probabilities. *Sankhyā Ser. A* **41** 184–197.

SPEED, T. P. and KIIVERI, H. (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist.* **14** 138–150.

TARJAN, R. E. and YANNAKAKIS, M. (1984). Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs and selectively reduce acyclic hypergraphs. *SIAM J. Comput.* **13** 566–579.

TATE, R. F. (1954). Correlation between a discrete and continuous variable. *Ann. Math. Statist.* **25** 603–607.

WERMUTH, N. (1980). Linear recursive equations, covariance selection and path analysis. *J. Amer. Statist. Assoc.* **75** 963–972.

WERMUTH, N. (1985). Data analysis and conditional independence structures. *Bull. Internat. Statist. Inst., Proc. 45th Session* **51-4** 24.2-1–24.2-14.

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

WERMUTH, N. and LAURITZEN, S. L. (1989). On substantive research hypotheses, conditional independence graphs and graphical chain models. *J. Roy. Statist. Soc. Ser. B* **51**. To appear.

WILLIAMS, D. A. (1976). Improved likelihood ratio tests for complete contingency tables. *Biometrika* **63** 33–37.

WOLD, H. O. A. (1954). Causality and econometrics. *Econometrica* **22** 162–177.

WRIGHT, S. (1921). Correlation and causation. *J. Agric. Res.* **20** 557–585.

WRIGHT, S. (1923). The theory of path coefficients: A reply to Niles' criticism. *Genetics* **8** 239–255.

WRIGHT, S. (1934). The method of path coefficients. *Ann. Math. Statist.* **5** 161–215.

INSTITUTE OF ELECTRONIC SYSTEMS
AALBORG UNIVERSITY CENTRE
STRANDVEJEN 19
DK-9000 AALBORG
DENMARK

PSYCHOLOGISCHES INSTITUT
JOHANNES GUTENBERG-UNIVERSITÄT
POSTFACH 3980
D-6500 MAINZ
FEDERAL REPUBLIC OF GERMANY