Generating Mixed Hierarchical Interaction Models by Selection

Steffen L. Lauritzen[1]

ABSTRACT: This note is concerned with the class of hierarchical interaction models for mixed discrete and continuous variables as defined by Edwards (1990) and modified by Lauritzen (1996). In particular it is shown that any hierarchical log-linear interaction model can be generated by selection on a set of response variables in a directed Markov model over what we have termed the *selection graph* of the model.
An inequality is established for the entries in the concentration matrix of any Gaussian undirected Markov distribution obtained by conditioning on the values of the response variables in the selection graph, thus demonstrating that not all such distributions can be generated in this way.
Finally it is shown that in the mixed case only hierarchical models of the type defined by Edwards (1990) can be generated by selection as above.

KEYWORDS: Bayesian networks; Conditional Gaussian distribution; Covariance selection; Gaussian graphical models; Log-linear interaction models; Recursive models.

1 Introduction

Although the class of log-linear models for contingency tables are well established in the statistical literature (Bishop, Fienberg and Holland, 1975), they remain less readily interpreted than undirected graphical models (Darroch, Lauritzen and Speed, 1980) or recursive graphical models (Wermuth and Lauritzen, 1983) also known as Bayesian networks (Pearl 1986, Jensen 1996).
In the following we shall investigate the possibility of interpreting hierarchical models through a selection process in simple recursive models. We begin by describing the basic elements of Conditional Gaussian (CG) distributions in Section 2 and proceed to review the basic concepts associated with mixed hierarchical interaction models.
We will show that all hierarchical log-linear models in the discrete case can

[1]Department of Mathematics, Aalborg University, Denmark.

indeed be generated by a simple selection process whereas this is not so for the Gaussian and mixed cases.

2 Conditional Gaussian Distributions

These distributions were introduced by Lauritzen and Wermuth (1984, 1989). Following Lauritzen (1996) we consider a finite set of variables V, partitioned as $V = \Delta \cup \Gamma$ into variables of *discrete* (Δ) and *continuous* (Γ) type. A typical element of the joint state space is denoted as in one of the possibilities below:

$$x = (x_\alpha)_{\alpha \in V} = (i, y) = \{(i_\delta)_{\delta \in \Delta}, (y_\gamma)_{\gamma \in \Gamma}\},$$

where i_δ are qualitative and y_γ are real-valued. A particular combination $i = (i_\delta)_{\delta \in \Delta}$ is referred to as a *cell* and the set of cells is denoted by $\mathcal{I}$. The joint distribution of the variables is supposed to have a density f with

$$\log f(x) = \log f(i, y) = g(i) + h(i)^\top y - y^\top K(i) y/2, \tag{1.1}$$

in which case we say that X follows a *Conditional Gaussian distribution* or, briefly, *CG distribution.* For each i, $g(i)$ is a real number, $h(i) = \{h(i)_\gamma\}_{\gamma \in \Gamma}$ a vector in $\mathcal{R}^{|\Gamma|}$, and $K(i) = \{k(i)_{\gamma\mu}\}_{\gamma,\mu \in \Gamma}$ a positive definite $|\Gamma| \times |\Gamma|$ matrix. The triple (g, h, K) is the *canonical characteristics* of the distribution. That X follows a CG distribution is equivalent to the statements

$$p(i) = P(X_\Delta = i) > 0 \quad \text{and} \quad \mathcal{L}(X_\Gamma \mid X_\Delta = i) = \mathcal{N}_{|\Gamma|}\{\xi(i), \Sigma(i)\},$$

where $X_A = (X_\alpha)_{\alpha \in A}$ and so on, $\mathcal{N}_{|\Gamma|}\{\xi(i), \Sigma(i)\}$ is the $|\Gamma|$-variate Gaussian distribution with mean $\xi(i)$ and covariance matrix Σ, and

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

The triple (p, ξ, Σ) are the *moment characteristics* of the distribution. The moment characteristics can be calculated from the canonical characteristics and *vice versa* by simple formulae, for details see Lauritzen (1996).
We say that the distribution is *homogeneous* if the covariance is independent of i, i.e. if $\Sigma(i) \equiv \Sigma$ or, equivalently, if $K(i) \equiv K$.

3 Interaction Expansion of CG Densities

To discuss interactions and interaction models we must expand the logarithm of the density into interaction terms. This is done in the following way:

$$\log f(i, y) = \sum_{d \subseteq \Delta} \lambda_d(i) + \sum_{d \subseteq \Delta} \sum_{\gamma \in \Gamma} \eta_d(i)_\gamma y_\gamma - \frac{1}{2} \sum_{d \subseteq \Delta} \sum_{\gamma,\mu \in \Gamma} \psi_d(i)_{\gamma\mu} y_\gamma y_\mu. \tag{1.2}$$

Here and in the following $\phi_d(i)$ for $d \subseteq \Delta$ denotes a function of $i = (i_\delta)_{\delta\in\Delta}$ that depends on i through $i_d = (i_\delta)_{\delta\in d}$ only.
We refer to the terms $\lambda_d(i)$ as *discrete interactions* among the variables in d. If $|d| = 1$ we also use the term *main effect* of the variable in d. The terms $\eta_d(i)_\gamma$ are *linear interactions* between γ and the variables in d. If $d = \emptyset$ we also use the term *linear main effect* of the variable γ. Finally the terms $\psi_d(i)_{\gamma\mu}$ are *quadratic interactions* between γ, μ and the variables in d. Again we speak of *quadratic main effects* if $\gamma = \mu$ and $d = \emptyset$.
The term $\lambda_\emptyset$ is constant and determined by normalization. If all variables are discrete and $\Gamma = \emptyset$, the discrete interactions are clearly the only ones present. Similarly, if all variables are Gaussian and $\Delta = \emptyset$ the only terms present in the expansion are the linear and quadratic main effects and quadratic interactions among pairs of continuous variables. For $d \neq \emptyset$, η_d and Ψ_d have no analogue in the pure cases. These terms are therefore proper *mixed interactions* and describe the interaction between the continuous and discrete variables. A homogeneous CG distribution has no mixed quadratic interactions, i.e. $\Psi_d = 0$ for $d \neq \emptyset$.
The interactions are related to the Markov property on an undirected graph $\mathcal{G} = (V, E)$ in the usual way, stated precisely in the following version of the Hammersley–Clifford theorem.

Theorem 1 *A CG distribution is Markov on the graph $\mathcal{G}$ if and only if the density has an expansion into interaction terms which satisfy*

$$\begin{aligned} \lambda_d(i) &= 0 \quad \textit{unless } d \textit{ is complete in } \mathcal{G} \\ \eta_d(i)_\gamma &= 0 \quad \textit{unless } d \cup \{\gamma\} \textit{ is complete in } \mathcal{G} \\ \psi_d(i)_{\gamma\mu} &= 0 \quad \textit{unless } d \cup \{\gamma, \mu\} \textit{ is complete in } \mathcal{G}. \end{aligned}$$

Proof: A simple proof is given in Lauritzen (1996), based on a general version of the Hammersley–Clifford theorem. The main elements of the proof are sketched below.
It is obvious that if the mentioned interaction terms vanish, then the density factorizes on the graph.
To see the converse we choose a reference configuration $x^* = (i^*, 0)$ and define interaction terms for $d \subseteq \Delta$ and $c \subseteq \Gamma$ as

$$\begin{aligned} \phi_{d\cup c}(x) &= \sum_{a\,:\,a\subseteq d\cup c} (-1)^{|(d\cup c)\setminus a|} \log f(x_a, x^*_{V\setminus a}) \\ &= \sum_{a\,:\,a\subseteq d} \sum_{b\,:\,b\subseteq c} (-1)^{|d\setminus a|+|c\setminus b|} \log f(i_a, i^*_{\Delta\setminus a}, y_b, 0_{\Gamma\setminus b}). \end{aligned}$$

Choosing the CG interactions appropriately leads to the following relations for $d \subseteq \Delta$, $c \subseteq \Gamma$, and $\gamma, \mu \in \Gamma$ with $\gamma \neq \mu$:

$$\begin{aligned} \phi_d(x) &= \lambda_d(i) \\ \phi_{d\cup\{\gamma\}}(x) &= \eta_d(i)_\gamma y_\gamma - \psi_d(i)_{\gamma\gamma} y_\gamma^2/2 \end{aligned}$$

$$\begin{aligned} \phi_{d\cup\{\gamma,\mu\}}(x) &= -\psi_d(i)_{\gamma\mu} y_\gamma y_\mu \\ \phi_{d\cup c}(x) &= 0 \text{ for } |c| > 2. \end{aligned}$$

Using conditional independence we find that the left-hand sides of these equations must be identically equal to zero unless the corresponding subsets are complete. The conclusions of the theorem follow. □

This result leads to the class of graphical interaction models, obtained by restricting interactions between a group of variables to be zero unless all variables in the group are mutual neighbours in the graph $\mathcal{G}$.

4 Hierarchical Interaction Models

In this section we consider more general types of restriction on the interaction structure than those determined from the Markov properties of a graph as above. On the other hand, we would not want to allow completely arbitrary patterns of restriction.
Following up on the discussion of Edwards (1990), Lauritzen (1996) defines a *hierarchical mixed interaction model* as a system of zero restrictions on the interaction parameters that satisfies the following rules

1. if λ_d is restricted to zero and $d \subseteq c$, then λ_c, η_c and ψ_c are all restricted to zero;
2. if $\eta_d()_\gamma$ is restricted to zero and $d \subseteq c$, then $\eta_c()_\gamma$ and $\psi_c()_{\gamma\mu}$ are both restricted to zero for all $\mu \in \Gamma$;
3. if $\psi_d()_{\gamma\mu}$ is restricted to zero and $d \subseteq c$, then $\psi_c()_{\gamma\mu}$ is restricted to zero.

Note that this defines a hierarchical interaction model in a way which is different from that of Edwards (1990) who also demands:

4. If $\psi_d()_{\gamma\gamma}$ is restricted to be zero, then so is $\psi_d()_{\gamma\mu}$ be for all μ.

We shall refer to models given by restrictions that satisfy 1 through 4 as *MIM models,* because the program MIM (Edwards, 1995) is based upon such models. However, as noted in the discussion of Edwards (1990) there seems to be no immediate justification for the conditon 4 above and indeed there are sensible models that are not MIM models, but are hierarchical models in the first sense. An example is the model which assumes variance homogeneity in the regression of Y on Z for different levels of a discrete variable I, see Example 6.31 of Lauritzen (1996) for details.

5 Invariance Properties of Hierarchical Models

In the following we show that hierarchical models satisfy natural invariance properties.
Suppose that we would demand the discrete interaction term λ_d to be identically equal to zero, but for some c with $d \subset c$ allow λ_c to vary freely. Then this would not restrict the class of densities at all, because if the density has been expanded as above in (1.2), we can rewrite the density in a way that fulfils the restriction by redefining

$$\tilde{\lambda}_c(i) = \lambda_c(i) + \lambda_d(i), \quad \tilde{\lambda}_d(i) = 0.$$

To avoid this happening, we demand of our model restriction that if λ_d is restricted to zero and $d \subseteq c$, then λ_c is also restricted to be zero. A similar property must hold for any of the other terms $\eta_d()_\gamma$ and $\psi_d()_{\gamma\mu}$ for the same reason.
Next we investigate the effect of a change in location and scale on the interaction terms. Let $Y_\gamma = a_\gamma \tilde{Y}_\gamma - b_\gamma$ for $a, b \in \mathcal{R}^{|\Gamma|}$. Then the density of $\tilde{X} = (I, \tilde{Y})$ can be found to be

$$\log \tilde{f}(i, \tilde{y}) = \tilde{g}(i) + \tilde{h}(i)^\top \tilde{y} - \frac{1}{2} \tilde{y}^\top \tilde{K}(i) \tilde{y},$$

where

$$\begin{aligned}
\tilde{g}(i) &= g(i) - h(i)^\top b - b^\top K(i) b/2 + \sum_\gamma \log |a_\gamma| \\
\tilde{h}(i)_\gamma &= a_\gamma h(i)_\gamma + a_\gamma \sum_{\mu \in \Gamma} k(i)_{\gamma\mu} b_\mu \\
\tilde{k}(i)_{\gamma\mu} &= a_\gamma a_\mu k(i)_{\gamma\mu}.
\end{aligned}$$

From these equations we derive similar expressions for the interactions:

$$\tilde{\lambda}_d(i) = \lambda_d(i) - \eta_d^\top b - b^\top \Psi_d b/2 + \sum_\gamma \log |a_\gamma| \tag{1.3}$$

$$\tilde{\eta}_d(i)_\gamma = a_\gamma \eta_d(i)_\gamma + a_\gamma \sum_{\mu \in \Gamma} \psi_d(i)_{\gamma\mu} b_\mu \tag{1.4}$$

$$\tilde{\psi}_d(i)_{\gamma\mu} = a_\gamma a_\mu \psi_d(i)_{\gamma\mu}. \tag{1.5}$$

Inspection of (1.3) shows that unless zero restrictions on the interaction terms are to be affected by a change of location, we must demand that whenever λ_d is restricted to be zero, then so are η_d and ψ_d. Further, (1.4) shows that if $\eta_d()_\gamma$ is restricted to zero, then so must $\psi_d()_{\gamma\mu}$ be for all $\mu \in \Gamma$. Finally (1.5) gives no restrictions on the interaction terms. On the other hand, the identities also show that a change of scale has no further effect. A change of scale affects the linear and quadratic interaction terms,

but it does not change absence to presence or conversely, as long as we have not restricted any interactions to have fixed values (other than zero). Thus we have given an argument that leads to restricting the interaction terms exactly as summarized in 1–3 of Section 4.
An alternative justification is to investigate the behaviour of the interaction terms under the process of *selection* or conditioning on a specific value of a variable. The conditional density of the remaining variables given that a variable X_α has a specific value x^*_α is obtained from the joint density by renormalization. This means that when $\alpha = \delta$ is discrete the interaction terms change as follows for $\delta \notin d$:

$$\begin{aligned}
\lambda^*_d(i) &= \lambda^*_d(i) + \lambda_{d\cup\{\delta\}}(i_d, i^*_\delta) \\
\eta^*_d(i)_\gamma &= \eta^*_d(i)_\gamma + \eta_{d\cup\{\delta\}}(i_d, i^*_\delta)_\gamma \\
\psi^*_d(i)_{\gamma\mu} &= \psi^*_d(i)_{\gamma\mu} + \psi_{d\cup\{\delta\}}(i_d, i^*_\delta)_{\gamma\mu}.
\end{aligned}$$

The interaction terms for those d which have $\delta \in d$ vanish and all other interaction terms are unchanged. Thus the conditions in 1 and 3 of Section 4 ensure that the model is robust under conditioning with a discrete variable. Similarly, when conditioning with the value y^*_μ of a continuous variable μ, the interaction terms change as

$$\begin{aligned}
\lambda^*_d(i) &= \lambda^*_d(i) + \eta_d(i)_\mu y^*_\mu - \psi_d(i)_{\mu\mu}(y^*_\mu)^2/2 \\
\eta^*_d(i)_\gamma &= \eta^*_d(i)_\gamma + \psi_d(i)_{\gamma\mu} y^*_\mu,
\end{aligned}$$

the interaction terms involving μ vanish, and the remaining terms are unchanged. So the conditions in 1 and 2 of Section 4 ensure that the model is robust under conditioning with a continuous variable.
For a thorough study of invariance properties of discrete hierarchical interaction models, see McCullagh (1997).

6 Generating Hierarchical Interaction Models

The previous considerations indicate that it can be difficult to interprete models determined by restricting interactions in a way that does not satisfy the conditions 1–3 of Section 4. However, this does not in itself lead to simple interpretation of a hierarchical interaction model. The most readily interpretable graphical models are the *recursive models*, i.e. the Markov models associated with a directed acyclic graph (DAG) as such a graph can be thought of as representing a generating mechanism (Cox and Wermuth, 1996) or it can be associated with a causal theory among the variables (Pearl, 1995).
Below we shall investigate to what extent hierarchical interaction models can be generated from simple recursive models by *selection* on one or more response variables.

6.1 The Discrete Case

Let us first investigate the discrete case. A log-linear hierarchical interaction model is determined by a generating class $\mathcal{A}$ of subsets of the set of variables Δ. Assume now that the logarithm of the probability of a given cell i is expanded into interaction terms as

$$\log p(i) = \sum_{d \in \mathcal{A}} \lambda_d(i).$$

Associate a binary random variable Z^d with every $d \in \mathcal{A}$ and define a DAG $\mathcal{D}$ over the variables $\tilde{\Delta} = \Delta \cup \mathcal{A}$ by having an arrow from $\delta \in \Delta$ to $d \in \mathcal{A}$ if and only if $\delta \in d$, and no further arrows. We shall term this DAG the *selection graph* for the hierarchical model. The selection graph of the hierarchical model with three variables $\Delta = \{1, 2, 3\}$ and only pairwise interactions $\mathcal{A} = \{\{1, 2\}, \{2, 3\}, \{1, 3\}\}$ is shown in Figure 1.

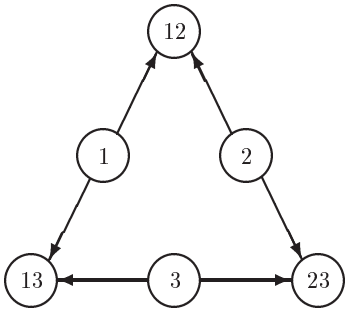

FIGURE 1. The selection graph of the hierarchical model with three variables and only pairwise interactions. Conditionally on the values of the variables at 12, 13, and 23, the interaction structure of the remaining variables only involves pairwise terms.

Note that the selection graph of an undirected graphical model has a response variable for each clique of the graph.
Next, let

$$c = \sup_{i \in \mathcal{I}, d \in \mathcal{A}} \lambda_d(i)$$

and define the probability distribution q as a directed Markov distribution over the selection graph with

$$q(Z^d = 1 \mid i) = 1 - q(Z_d = 0 \mid i) = \exp\{\lambda_d(i) - 2c\}, \quad q(i) = 1/|\mathcal{I}|.$$

This clearly defines a probability distribution which satisfies the directed

Markov property over the selection graph and we have

$$q(i \mid Z^d = 1 \text{ for all } d \in \mathcal{A}) \propto |\mathcal{I}|^{-1} \prod_{d \in \mathcal{A}} \exp\{\lambda_d(i) - 2c\} \propto p(i)$$

and thus we must have

$$q(i \mid Z^d = 1 \text{ for all } d \in \mathcal{A}) = p(i).$$

On the other hand, assume that q is strictly positive and directed Markov over the selection graph, i.e.

$$q(i, z) = \prod_{\delta \in \Delta} q(i_\delta) \prod_{d \in \mathcal{A}} q(z^d \mid i_d).$$

Let also z_0^d be a fixed value of z^d. Then the conditional distribution of the variables in Δ given $Z^d = z_0^d$ for all $d \in \mathcal{A}$ satisfies the restrictions of the hierarchical model with generating class $\mathcal{A}$. This is easily seen by associating with each δ a specific $d(\delta)$ with $\delta \in d(\delta)$ and defining

$$\lambda_d(i) = \log q(z_0^d \mid i_d) + \sum_{\delta : d(\delta) = d} \log q(i_\delta).$$

We summarize the findings above in the following theorem:

Theorem 2 *A probability distribution over a set Δ of discrete random variables satisfies the restrictions of a hierarchical log-linear interaction model with generating class $\mathcal{A}$ if and only if it can be obtained by conditioning on the values of the response variables in a recursive model over the selection graph of $\mathcal{A}$.*

6.2 The Gaussian Case

In the case where all variables are continuous and their joint distribution is multivariate Gaussian, there is no difference between the class of hierarchical and the class of graphical models, as all interaction terms involve only pairs of variables. A given undirected graph $\mathcal{G} = (\Gamma, E)$ determines a Gaussian graphical model by assuming that $Y = Y_\Gamma$ has a multivariate normal distribution $\mathcal{N}_{|\Gamma|}(0, \Sigma)$ where $K = \Sigma^{-1}$ is an element of $\mathcal{S}^+(\mathcal{G})$, the set of positive definite matrices K with $k_{\gamma\mu} = 0$ unless $\gamma \sim \mu$ in $\mathcal{G}$. Equivalently, Y_γ and Y_μ are assumed to be conditionally independent unless $\gamma \sim \mu$ in $\mathcal{G}$. An undirected graphical Gaussian model is also known as a *covariance selection model* (Dempster, 1972).
In analogy of the discrete case we shall investigate to which extent an arbitrary covariance selection model can be obtained from a recursive Gaussian graphical model by selection, i.e. by conditioning on particular values of the response variables.

As before we associate a *selection graph* with the graph $\mathcal{G}$. The selection graph $\mathcal{D}$ has vertices $\tilde{\Gamma} = \Gamma \cup E$ where E is the set of edges in $\mathcal{G}$. There is an arrow from γ to $e \in E$ in $\mathcal{D}$ if and only if γ is an endpoint of the edge e. Figure 2 shows the selection graph of the 4-cycle $\mathcal{G} = (\{1,2,3,4\}, \{\{1,2\},\{2,3\},\{3,4\},\{1,4\}\})$. Clearly, the conditional dis-

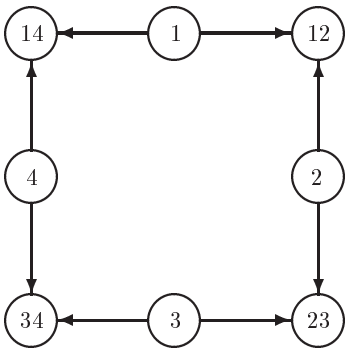

FIGURE 2. The selection graph of the covariance selection model corresponding to the 4-cycle. The variables Y_1 and Y_3 are conditionally independent given Y_2, Y_4, and the four response variables. Similarly, Y_2 and Y_4 are conditionally independent given Y_1, Y_3, and the four response variables.

tribution of the original variables given any value of the response variables in the selection graph will satisfy the Markov restrictions of the original graph. However, the converse to this assertion is not true, as we shall see below. In fact, the concentration matrix of any distribution obtainable by conditioning on the response variables satisfies an additional inequality.

Theorem 3 *If there exists a multivariate Gaussian directed Markov distribution on the selection graph $\mathcal{D}$ such that the conditional distribution $\mathcal{L}(Y_\Gamma \,|\, Y_E = y_E^*)$ is $\mathcal{N}_{|\Gamma|}(0, K^{-1})$, then K is in $\mathcal{S}^+(\mathcal{G})$. Further K satisfies the inequality*

$$\mathrm{tr}(K) \geq 2 \sum_{\{\gamma,\mu\}:\{\gamma,\mu\}\in E} |k_{\gamma\mu}|. \tag{1.6}$$

Proof: Assume that $Y_{\Gamma\cup E}$ follows a multivariate Gaussian and directed Markov distribution on the selection graph and let Λ denote the diagonal concentration matrix of Y_Γ. Let further $e = \{\gamma, \mu\} \in E$ and assume

$$\mathcal{L}(Y_e \,|\, Y_\Gamma) = \mathcal{N}(\alpha_\mu^\gamma y_\gamma + \alpha_\gamma^\mu y_\mu + c_{\gamma\mu}, \sigma_{\gamma\mu}^2).$$

Define $a_\mu^\gamma = \alpha_\mu^\gamma / \sigma_{\gamma\mu}$ and $a_\gamma^\mu = \alpha_\gamma^\mu / \sigma_{\gamma\mu}$. The concentration matrix K of the conditional distribution of Y_Γ given any value of Y_E is found to have

elements

$$k_{\gamma\gamma} = \lambda_{\gamma\gamma} + \sum_{\mu:\mu\sim\gamma} (a_\mu^\gamma)^2, \quad k_{\gamma\mu} = a_\mu^\gamma a_\gamma^\mu.$$

But as we have the elementary inequality

$$(a_\mu^\gamma)^2 + (a_\gamma^\mu)^2 \geq 2|a_\mu^\gamma a_\gamma^\mu| = 2|k_{\gamma\mu}|,$$

we find

$$\begin{aligned} \mathrm{tr}(K) &= \sum_{\gamma\in\Gamma} k_{\gamma\gamma} = \sum_{\gamma\in\Gamma} \left\{ \lambda_{\gamma\gamma} + \sum_{\mu:\mu\sim\gamma} (a_\mu^\gamma)^2 \right\} \\ &= \mathrm{tr}(\Lambda) + \sum_{\{\gamma,\mu\}:\{\gamma,\mu\}\in E} \left\{ (a_\mu^\gamma)^2 + (a_\gamma^\mu)^2 \right\} \\ &\geq 2 \sum_{\{\gamma,\mu\}:\{\gamma,\mu\}\in E} |k_{\gamma\mu}| \end{aligned}$$

which was to be proved. □

Note that for $|\Gamma| = 2$ the inequality (1.6) is automatically fulfilled. In fact, from the positive definiteness of K we get for all pairs $\{\gamma, \mu\}$ that

$$2|k_{\gamma\mu}| \leq 2\sqrt{k_{\gamma\gamma}k_{\mu\mu}} \leq k_{\gamma\gamma} + k_{\mu\mu}.$$

An example of a concentration matrix which is not obtainable by selection on the response variables is the 3×3-matrix with 1 in the diagonal and ρ outside the diagonal with $\rho > 2/3$. Because then $\mathrm{tr}(K) = 3$ but the right-hand side of the inequality (1.6) amounts to 6ρ which is then surely greater than 4.
Similarly, let K be the 4×4-matrix with diagonal elements equal to 1 and the off-diagonal elements given as

$$k_{12} = \rho, \quad k_{23} = \rho, \quad k_{34} = \rho, \quad k_{14} = -\rho,$$

and all other elements equal to zero. Then K satisfies the restrictions of the 4-cycle and K is positive definite if and only if $|\rho|^2 < 1/2$. For $\rho = 0.7$, the right-hand side of the inequality (1.6) becomes equal to $8|\rho| = 5.6$, and as $\mathrm{tr}(K) = 4$, this concentration matrix cannot be obtained by conditioning on the response variables in the selection graph of Figure 2.

6.3 The Mixed Case

We finally consider the mixed case and try to investigate what type of restrictions we can obtain on the conditional distribution of a system of discrete and continuous variables, given a set of responses. Clearly, the situation here is more complicated than in any of the pure cases and it is

therefore hardly likely that a simple and complete description can be given of the model restrictions obtainable by conditioning on a system of response variables. Clearly a discrete term $\lambda_d(i)$ in the interaction expansion (1.2) can be generated by conditioning on a binary response variable Z_d as in the discrete case. But consider a continuous response variable Z and assume that

$$\mathcal{L}(Z \mid i, y) = \mathcal{N}\{\alpha(i) + \beta(i)^\top y, \sigma^2(i)\}$$

with the density of $X = (I, Y)$ being $f(i, y)$. Let

$$a(i) = \{z_0 - \alpha(i)\}/\sigma(i), \quad b(i) = \beta(i)/\sigma(i).$$

Then conditioning on $Z = z_0$ yields the density

$$\begin{aligned} \log f(i, y \mid z_0) &= \text{constant} + \log f(i, y) \\ &\quad -a(i)^2/2 + a(i)b(i)^\top y - \{y^\top b(i)\}^2/2 \\ &= \text{constant} + \log f(i, y) + \lambda(i) + \eta(i)^\top y - y^\top \Psi(i) y/2, \end{aligned}$$

where

$$\lambda(i) = -a(i)^2/2, \quad \eta(i) = a(i)b(i), \quad \Psi(i) = b(i)b(i)^\top.$$

If the distribution of Z is only allowed to depend on a subset $d \cup c$ of variables where $d \subseteq \Delta$ and $c \subseteq \Gamma$, this results in each of these functions only depending on i through i_d and further that

$$b(i)_\gamma = b_d(i)_\gamma = 0 \text{ unless } \gamma \in c.$$

Thus the conditioning on Z creates interaction terms of the form

$$\lambda_d(i), \quad \eta_d(i)_\gamma, \quad \psi_d(i)_{\gamma\mu}, \quad \gamma, \mu \in c.$$

If both $\beta(i)$ and $\sigma^2(i)$ are assumed to be independent of i, then no mixed quadratic interaction terms are created.
In the case where the regression of Z on Y_γ is assumed to be homogeneous and with slope independent of i, i.e.

$$\mathcal{L}(Z \mid i, y) = \mathcal{N}\{\alpha(i) + \beta^\top y, \sigma^2\},$$

with $\alpha(i)$ only depending on i through i_d, we get interaction terms of the form

$$\lambda_d(i) = -a(i)^2/2, \quad \eta_d(i)_\gamma = a(i)b_\gamma, \quad \psi_{\gamma\mu} = b_\gamma b_\mu$$

but no mixed quadratic interactions.
Note that whenever a term of the form $\psi_d(i)_{\gamma\mu}$ is created in any of the above ways, there is simultaneously created terms of the form $\psi_d(i)_{\gamma\gamma}$, $\psi_d(i)_{\mu\mu}$, $\eta_d(i)_\gamma$, and $\eta_d(i)_\mu$. Thus models obtained by selection in this way will satisfy condition 4 of Section 4 and thus satisfy the restrictions of a MIM model. So only models of the MIM type have a chance of being generated by conditioning on response variables. Additionally, certain inequalities will also restrict the class of models that can be generated in this way.

References

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, Massachusetts.

Cox, D. R. and Wermuth, N. (1996). *Multivariate Dependencies*. Chapman and Hall, London.

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

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

Edwards, D. (1990). Hierarchical interaction models (with discussion). *Journal of the Royal Statistical Society, Series B*, **52**, 3–20 and 51–72.

Edwards, D. (1995). *Introduction to Graphical Modelling*. Springer-Verlag, New York.

Jensen, F. V. (1996). *An Introduction to Bayesian Networks*. University College London Press, London.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford

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

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

McCullagh, P. (1997). Invariance and linear models. Technical Report No. 450. Department of Statistics, University of Chicago.

Pearl, J. (1986). Fusion, propagation and structuring in belief networks. *Artificial Intelligence*, **29**, 241–288.

Pearl, J. (1995). Causal diagrams for empirical research (with discussion). *Biometrika*, **82**, 669–710.

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