

# Collapsibility of Graphical CG–Regression Models

Running title: Collapsibility of CG–Regressions

VANESSA DIDELEZ

*University College London*

DAVID EDWARDS

*Novo Nordisk*

ABSTRACT: CG–regressions are multivariate regression models for mixed continuous and discrete responses that result from conditioning in the class of conditional Gaussian (CG) models. Their conditional independence structure can be read off a marked graph. The property of collapsibility, in this context, means that the multivariate CG–regression can be decomposed in lower dimensional regressions that are still CG and are consistent with the corresponding subgraphs. We derive conditions for this property that can easily be checked on the graph, and indicate computational advantages of this kind of collapsibility. Further, a simple graphical condition is given for checking whether a decomposition into univariate regressions is possible.

*Keywords:* chain graphs, conditional independence graphs, decomposability, hierarchical models, TM algorithm

## 1 Introduction

We study in this paper a key property of a class of models used in graphical modelling with mixed discrete and continuous data, the CG–regression models. These describe the distribution of a set of response variables given a set of covariates, where both sets may be comprised of discrete and continuous variables. Polytomous logistic regression is the special case in which there is one discrete response only.

The CG–regression models arise by conditioning in a class of joint models based

on the conditional Gaussian (CG) distribution. These CG–distribution models form the basis for graphical modelling using undirected graphs (see Lauritzen, 1996; Edwards, 2000). They are of two types. The graphical interaction models (Lauritzen and Wermuth, 1989; Lauritzen, 1996) restrict the CG–distributions so as to satisfy a set of conditional independence constraints corresponding to an undirected graph. The hierarchical interaction models (Edwards, 1990 and 2000) extend the graphical interaction models by permitting further parametric constraints. In this paper we study the conditional models derived from the former model class, and we refer to them as graphical CG–regression models.

Apart from their general utility as regression models for multivariate mixed, that is to say, discrete and continuous response variables, the CG–regression models play an important rôle as building blocks for chain graph models (Lauritzen, 1996, section 6.5). The models may be fitted using the TM algorithm (Edwards and Lauritzen, 2001), as we describe briefly in § 4.

The property of collapsibility in the sense used here is due to Lauritzen (1989); see also Asmussen and Edwards (1983) and Frydenberg (1990b). Roughly speaking, when a model is collapsible onto a variable subset, it can be decomposed into two simpler models, a marginal model on the subset, and a conditional model describing the dependence of the remaining variables on the subset variables. When available, this simplification leads to a deeper understanding of the model and can have important implications for estimation and hypothesis testing.

The structure of the paper is as follows. In § 2 we introduce some notation, describe the model families in more detail, and define collapsibility. In § 3 we study the collapsibility of graphical CG–regression models. The main result of the paper is Theorem 2, which gives necessary and sufficient conditions for collapsibility for these models. The implications of Theorem 2, in particular for maximum likelihood estimation, are presented in § 4. In § 5, Theorem 3 characterises models that can be decomposed into a sequence of univariate CG–regressions. In § 6 we indicate further open questions.

## 2 Basic notions

We consider classes of models for a multivariate random vector  $X = X_{\mathcal{V}}$  with index set  $\mathcal{V} = \Delta \cup \Gamma$ , where  $\Delta$  is a set of discrete variables and  $\Gamma$  a set of continuous variables. We also use the notation  $I = X_{\Delta}$  for the discrete and  $Y = X_{\Gamma}$  for the continuous components. Further, for  $a \subset \mathcal{V}$  we define  $I_a = I_{a \cap \Delta}$  and  $Y_a = Y_{a \cap \Gamma}$ . We use  $f$  as generic symbol for a density function, and  $f_a$  and  $f_{a|c}$  (where  $a, c \subset \mathcal{V}$ ) to denote marginal and conditional densities, respectively.

### 2.1 CG–distributions and CG–regressions

We say that  $X = (I, Y)$  is CG–distributed when  $I$  is multinomially distributed and the conditional density of  $Y$  given  $I = i$  is Gaussian (Lauritzen and Wermuth, 1989). The joint density takes the form

$$f(x) = f(i, y) = \exp\{\alpha(i) + \beta(i)'y - \frac{1}{2}y'\Omega(i)y\} \quad (1)$$

where, for all  $i$ ,  $\alpha(i)$  is a real number,  $\beta(i)$  is a  $|\Gamma|$ –vector, and  $\Omega(i)$  is a symmetric positive-definite  $|\Gamma| \times |\Gamma|$ –matrix. These are known as the discrete, linear and quadratic canonical parameters.

Alternatively, the moment parametrisation may be adopted. This uses  $p(i)$ , the cell probabilities for the discrete variables, and  $\xi(i)$  and  $\Sigma(i)$ , the mean vector and covariance matrix of the continuous given the discrete variables. The relation between the canonical and moment parameterisations can be found in Lauritzen (1996, p.159). The former gives simpler expressions for conditional distributions whereas the latter is more suitable for marginalising. Since we consider both operations we use both representations.

The homogeneous CG–distributions are of the above form but restrict the quadratic parameters to be constant over  $i$ . That is to say,  $\Omega(i) = \Omega$  or, equivalently,  $\Sigma(i) = \Sigma$ , for all  $i$ . For the present purpose we do not need to treat the homogeneous distributions or models separately.

For  $c \subset \mathcal{V}$  and  $a = \mathcal{V} \setminus c$ , we say that  $X_a$  follows a CG–regression given  $X_c = x_c = (i_c, y_c)$ , if the conditional distribution of  $X_a$  given  $X_c = x_c = (i_c, y_c)$  is conditional Gaussian with moment parameters of the form

$$\begin{aligned} \log p(i_a|i_c, y_c) &= u(i_a|i_c) + v(i_a|i_c)^\top y_c - y_c^\top W(i_a|i_c) y_c \\ &\quad - \log \kappa(i_c, y_c) \end{aligned} \tag{2}$$

$$\xi(i_a|i_c, y_c) = c(i_a|i_c) + C(i_a|i_c) y_c \tag{3}$$

$$\Sigma(i_a|i_c, y_c) = D(i_a|i_c), \tag{4}$$

for some sextuple of parameters  $(u, v, W, c, C, D)$ . For all  $i_a$  and  $i_c$

$u(i_a|i_c)$  is a real number representing the interaction effects among the discrete components,

$v(i_a|i_c)$  is a  $|\Gamma \cap c|$ –vector representing the linear effects of the continuous components  $Y_c$  on the discrete ones,

$W(i_a|i_c)$  is a symmetric  $|\Gamma \cap c| \times |\Gamma \cap c|$ –matrix representing the quadratic effects of the continuous components  $Y_c$  on the discrete ones,

$c(i_a|i_c)$  is a  $|\Gamma \cap a|$ –vector representing the interaction effects of the discrete components on the mean of the continuous  $Y_a$ ,

$C(i_a|i_c)$  is a  $|\Gamma \cap a| \times |\Gamma \cap c|$ –matrix representing the linear effects of  $Y_c$  on  $Y_a$ ,

$D(i_a|i_c)$  is a symmetric and positive definite  $|\Gamma \cap a| \times |\Gamma \cap a|$ –matrix representing the interaction effects of the discrete components on the covariance structure of  $Y_a$ .

In (2),  $\kappa(i_c, y_c)$  is a normalising constant such that  $\sum_{i_a} p(i_a|i_c, y_c) = 1$ .

That conditioning on  $X_c = x_c$  in (1) leads to conditional distributions of this form was derived by Lauritzen and Wermuth (1989). They also showed that any distribution with the above form can be derived in this way, that is, for any conditional distribution  $f_{a|c}$  with the above parameterisation (2) – (4) there exists a joint CG–distribution  $f'_{a \cup c}$  such that  $f'_{a|c} = f_{a|c}$ .

## 2.2 Conditional independence restrictions

The models we consider in § 3 are generated by an undirected graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , that is they satisfy specific conditional independence restrictions implied by the graph. The vertices  $\mathcal{V}$  represent the elements of  $X_{\mathcal{V}}$ , the variables, whereas the edges  $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$  indicate conditional dependencies. More precisely, a distribution  $f$  is called  $\mathcal{G}$ -Markovian if for any three sets  $a, b, s \subseteq \mathcal{V}$ ,  $X_a$  is conditionally independent of  $X_b$  given  $X_s$  under  $f$  whenever  $s$  separates  $a$  and  $b$  in  $\mathcal{G}$ , that is when any path between  $a$  and  $b$  is intersected by  $s$ . We denote this conditional independence by  $a \perp\!\!\!\perp b \mid s$  (Dawid, 1979).

For example, the graph  $\mathcal{G}$  shown in Figure 1b has three vertices: two discrete ones,  $I$  and  $J$ , and one continuous,  $X$ . By convention the discrete variables are represented as filled circles and the continuous variables as hollow circles. The graph induces one conditional independence relation:  $I \perp\!\!\!\perp J \mid X$ , because  $X$  separates  $I$  and  $J$ , that is to say, all paths between  $I$  and  $J$  intersect  $X$ . For this  $\mathcal{G}$ , to say that a distribution  $f$  on  $(I, J, X)$  is  $\mathcal{G}$ -Markovian is to say that  $I \perp\!\!\!\perp J \mid X$  under  $f$ .

The graphical interaction model corresponding to a graph  $\mathcal{G}$ , which we write as  $\mathcal{M}(\mathcal{G})$ , consists precisely of the  $\mathcal{G}$ -Markovian CG-distributions. The conditional independencies implied by the graph  $\mathcal{G}$  induce specific parametric constraints on the joint distribution (1) (see, for example, Lauritzen, 1996, p.174). Obviously,  $\mathcal{M}(\mathcal{G})$  includes as special cases loglinear models if all variables are discrete and multivariate normal models if all variables are continuous. The conditional independence restrictions induced by a graph have been studied in detail for loglinear models by Darroch et al. (1980). Note that graphical loglinear models always include all higher order interactions except for those that vanish due to conditional independencies. Graphical Gaussian models are also known as covariance selection models (Dempster, 1972) as the conditional independence restrictions correspond to zeroes in the inverse covariance matrix.

Before continuing with conditional models, we need some further terminology. Vertices that are joined by an edge are called *adjacent*, and the *boundary*

$\text{bd}(a)$  of a set  $a \subset \mathcal{V}$  is defined to be the set of all vertices in  $\mathcal{V} \setminus a$  which are adjacent to some vertex in  $a$ . Further, a subgraph  $\mathcal{G}_a$ ,  $a \subset \mathcal{V}$ , is defined as  $\mathcal{G}_a = (a, \mathcal{E} \cap (a \times a))$ . A *connected component* of a graph is a subset of vertices  $a \subset \mathcal{V}$  such that any vertex in the subgraph  $\mathcal{G}_a$  is connected by some path to any other vertex. A subgraph is called *complete* if there is an edge between any pair of vertices in this subset.

We now turn to the conditional models. Given a set of covariates  $c \subset \mathcal{V}$  and a graph  $\mathcal{G}$ , we define the induced graphical CG–regression model as  $\mathcal{M}(\mathcal{G})^c = \{f_{a|c} : f \in \mathcal{M}(\mathcal{G})\}$ ,  $a = \mathcal{V} \setminus c$ . More strictly,  $\mathcal{M}(\mathcal{G})^c$  contains all conditional distributions such that there exists a realisation  $(i_c, y_c)$  of  $X_c = (I_c, Y_c)$  with  $f_{a|c}$  being the density of the conditional distribution of  $X_a = (I_a, Y_a)$  given  $(I_c, Y_c) = (i_c, y_c)$ , depending on the latter through (2) – (4). The class of CG–regressions  $\mathcal{M}(\mathcal{G})^c$  is thus parameterised according to (2) – (4) imposing, in addition, some constraints corresponding to the conditional independencies induced by the graph  $\mathcal{G}$ .

Clearly, the conditional distribution does not contain any information on the conditional independencies among the components of  $X_c$ , so for any  $f \in \mathcal{M}(\mathcal{G})$  the conditional distribution  $f_{a|c}$  does not depend on the structure of the subgraph  $\mathcal{G}_c$ . However, we may deduce statements like for example  $a \perp\!\!\!\perp c_1 | c_2$  for  $a \subseteq \mathcal{V} \setminus c$  and  $c_1, c_2 \subseteq c$ , if for instance  $f_{a|c_1, c_2}(x_a | x_{c_1 \cup c_2}) = f_{a|c_2}(x_a | x_{c_2})$ . A method for reading off a graph all the conditional independencies that are contained in the regression model thus has to modify the original graph  $\mathcal{G}$  by making the subgraph  $\mathcal{G}_c$  complete, so that no specific independencies among the variables in  $c$  are postulated. We denote the resulting graph by  $\mathcal{G}^c$  and call it the *conditioned* graph. The conditional independence restrictions for  $\mathcal{M}(\mathcal{G})^c$  are then given by

$$\forall a \subset \mathcal{V} \setminus c \text{ and } b, s \subset \mathcal{V}, a \perp\!\!\!\perp b | s \text{ whenever } s \text{ separates } a \text{ and } b \text{ in } \mathcal{G}^c. \quad (5)$$

We may therefore restate the definition of the  $\mathcal{G}$ –Markovian property for regression models as follows.

**Definition:** If a class of CG–regressions on  $X_{\mathcal{V}\setminus c}$  given  $X_c$ ,  $c \in \mathcal{V}$ , satisfies the conditional independencies (5) for a graph  $\mathcal{G}$  we say that it is  $\mathcal{G}^c$ –Markovian.

This allows us to characterise the distributions in  $\mathcal{M}(\mathcal{G})^c$  as precisely those conditional distributions resulting from the  $\mathcal{G}^c$ –Markovian CG–regressions, implying  $\mathcal{M}(\mathcal{G})^c = \mathcal{M}(\mathcal{G}^c)^c$ . A useful graphical representation for the conditional independence restrictions in  $\mathcal{M}(\mathcal{G})^c$  is given by modifying  $\mathcal{G}$  as follows: make the subgraph  $\mathcal{G}_c$  complete and replace all edges between  $\mathcal{G}_{\mathcal{V}\setminus c}$  and  $\mathcal{G}_c$  by directed ones pointing towards  $\mathcal{G}_{\mathcal{V}\setminus c}$ . The resulting graph is then a chain graph (Wermuth and Lauritzen, 1990) and the conditional independence properties (5) coincide with the chain graph Markov properties (Frydenberg, 1990a) except for those within  $\mathcal{G}_c$  which are not of interest for our purposes.

### 2.3 Collapsibility

Now let  $a, b, c$  be a partition of  $\mathcal{V}$ . We consider marginalising over  $b$  and conditioning on  $c$ . For example, for a density  $f = f_{a\cup b\cup c}$ , marginalising over  $b$  yields  $f_{a\cup c}$ , conditioning on  $c$  yields  $f_{a\cup b|c}$ , and both result in  $f_{a|c}$ . The class of all marginal distributions is defined as  $\mathcal{M}(\mathcal{G})_{a\cup c} = \{f_{a\cup c} : f \in \mathcal{M}(\mathcal{G})\}$ . Further, given a model  $\mathcal{M}(\mathcal{G})$ , we can marginalise to the model  $\mathcal{M}(\mathcal{G}_{a\cup c})$  induced by the subgraph  $\mathcal{G}_{a\cup c}$  or condition to  $\mathcal{M}(\mathcal{G})^c$ , or both to  $\mathcal{M}(\mathcal{G}_{a\cup c})^c$ .

Since the class of CG–distributions is not closed under marginalisation (Lauritzen and Wermuth, 1989), it is possible that  $\mathcal{M}(\mathcal{G}_{a\cup c}) \neq \mathcal{M}(\mathcal{G})_{a\cup c}$ ; the former,  $\mathcal{M}(\mathcal{G}_{a\cup c})$ , consists by definition of all CG–distributions on the subgraph  $\mathcal{G}_{a\cup c}$  while the distributions in the latter,  $\mathcal{M}(\mathcal{G})_{a\cup c}$ , may or may not be CG after marginalising the CG–distributions on  $\mathcal{G}$  over  $b$ . However, given any  $f_{a\cup c} \in \mathcal{M}(\mathcal{G}_{a\cup c})$  we can always construct an  $f' \in \mathcal{M}(\mathcal{G})$  such that  $f'_{a\cup c} = f_{a\cup c}$  by setting  $f' = f_b f_{a\cup c}$  for some  $f_b \in \mathcal{M}(\mathcal{G}_b)$ . The same is true for the regression models  $\mathcal{M}(\mathcal{G})^c$ . It follows that  $\mathcal{M}(\mathcal{G}_{a\cup c}) \subseteq \mathcal{M}(\mathcal{G})_{a\cup c}$  and  $\mathcal{M}(\mathcal{G}_{a\cup c})^c \subseteq \mathcal{M}(\mathcal{G})_{a\cup c}^c$ . The special case of collapsibility of the two types of models is given when the two sets coincide, respectively, as formulated in the next definition.

**Definitions:**

1. Let  $a \subseteq \mathcal{V}$ . If  $\mathcal{M}(\mathcal{G})_a = \mathcal{M}(\mathcal{G}_a)$ , that is, if  $f \in \mathcal{M}(\mathcal{G}) \Rightarrow f_a \in \mathcal{M}(\mathcal{G}_a)$ , we say that the graphical interaction model  $\mathcal{M}(\mathcal{G})$  is *collapsible* onto  $a$  and write this as  $\mathcal{M}(\mathcal{G}) \xrightarrow{\text{coll}} a$ .
2. Let  $a, b, c$  be a partition of  $\mathcal{V}$ . If  $\mathcal{M}(\mathcal{G})_{a \cup c}^c = \mathcal{M}(\mathcal{G}_{a \cup c})^c$ , that is, if  $f \in \mathcal{M}(\mathcal{G}) \Rightarrow f_{a|c} \in \mathcal{M}(\mathcal{G}_{a \cup c})^c$ , we say that the graphical CG–regression model  $\mathcal{M}(\mathcal{G})^c$  is *collapsible* onto  $a \cup c$  and write this as  $\mathcal{M}(\mathcal{G})^c \xrightarrow{\text{coll}} a \cup c$ .

Note that  $\mathcal{M}(\mathcal{G}) \xrightarrow{\text{coll}} a$  is equivalent to  $\mathcal{M}(\mathcal{G}) = \mathcal{M}(\mathcal{G}_a) \times \mathcal{M}(\mathcal{G})^a$ , that is any  $f \in \mathcal{M}(\mathcal{G})$  can be factorised as  $f = f_a f_{b|a}$ ,  $b = \mathcal{V} \setminus a$ , where  $f_a \in \mathcal{M}(\mathcal{G}_a)$ ,  $f_{b|a} \in \mathcal{M}(\mathcal{G})^a$ , and  $f_a$  and  $f_{b|a}$  are variation independent (Frydenberg, 1990b). Analogously we will show in Section 4 that  $\mathcal{M}(\mathcal{G})^c \xrightarrow{\text{coll}} a \cup c$  is equivalent to  $\mathcal{M}(\mathcal{G})^c = \mathcal{M}(\mathcal{G}_{a \cup c})^c \times \mathcal{M}(\mathcal{G})^{a \cup c}$ , so that there exist variation independent  $f_{a|c} \in \mathcal{M}(\mathcal{G}_{a \cup c})^c$  and  $f_{b|a \cup c} \in \mathcal{M}(\mathcal{G})^{a \cup c}$  such that  $f_{a \cup b|c} = f_{a|c} f_{b|a \cup c}$  for a partition  $a, b, c$  of  $\mathcal{V}$ .

Necessary and sufficient conditions for collapsibility of graphical interaction models were found by Frydenberg (1990b). Essentially, two conditions are involved. One ensures that the marginal distribution is again CG and the other ensures that marginalisation does not destroy conditional independences among the remaining variables. To state these conditions we need a few more definitions.

**Definitions:**

1. A subset  $b \subseteq \mathcal{V}$  is *simplicial* in  $\mathcal{G}$  if  $\text{bd}(b)$  is complete in  $\mathcal{G}$ .
2. A subset  $b \subseteq \mathcal{V}$  is *strong* in  $\mathcal{G}$  if  $b \cap \Delta \neq \emptyset \Rightarrow \text{bd}(b) \subseteq \Delta$ .
3. A subset  $b \subseteq \mathcal{V}$  is a *simplicial (strong) collection* in  $\mathcal{G}$  if every connected component of  $b$  is simplicial (strong) in  $\mathcal{G}$ .

The conditions for collapsibility in undirected CG–models proved by Frydenberg (1990b) are as follows:

**Theorem 1** *For a graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  and  $a \subseteq \mathcal{V}$ ,  $\mathcal{M}(\mathcal{G}) \xrightarrow{\text{coll}} a$  if and only if  $\mathcal{V} \setminus a$  is a strong and simplicial collection in  $\mathcal{G}$ .*

The aim of the next section is to find corresponding conditions for the collapsibility of graphical CG–regression models.

### 3 Collapsibility of graphical CG–regressions

Before considering collapsibility in the conditional models we derive some general results that are interesting in themselves and facilitate the derivation of the main result. The first of these states that the collapsibility property is associative.

**Lemma 1** *If  $\mathcal{M}(\mathcal{G}) \xrightarrow{\text{coll}} a \cup c$ , and  $\mathcal{M}(\mathcal{G}_{a \cup c}) \xrightarrow{\text{coll}} c$ , then  $\mathcal{M}(\mathcal{G}) \xrightarrow{\text{coll}} c$ .*

**Proof:** This follows immediately from the corresponding graphical marginalisation property, which implies that  $\mathcal{G}_c = \{\mathcal{G}_{a \cup c}\}_c$ . The result generalises Lemma 3.4 of Madigan and Mosurski (1990).

A non–obvious corollary is that collapsibility onto a subset implies that all derived marginal models are also collapsible onto the subset:

**Lemma 2** *If  $\mathcal{M}(\mathcal{G}) \xrightarrow{\text{coll}} c$  then  $\mathcal{M}(\mathcal{G}_{a \cup c}) \xrightarrow{\text{coll}} c$ .*

**Proof:** For any  $f_{a \cup c} \in \mathcal{M}(\mathcal{G}_{a \cup c})$ , there exists a  $f' \in \mathcal{M}(\mathcal{G})$  such that  $f'_{a \cup c} = f_{a \cup c}$ . Since  $\mathcal{M}(\mathcal{G}) \xrightarrow{\text{coll}} c$ ,  $f'_c \in \mathcal{M}(\mathcal{G}_c)$ , hence  $f_c = f'_c \in \mathcal{M}(\mathcal{G}_c)$  as required.

This allows Lemma 1 to be sharpened slightly as follows:

**Lemma 3** *If  $\mathcal{M}(\mathcal{G}) \xrightarrow{\text{coll}} a \cup c$ , then  $\mathcal{M}(\mathcal{G}_{a \cup c}) \xrightarrow{\text{coll}} c$  if and only if  $\mathcal{M}(\mathcal{G}) \xrightarrow{\text{coll}} c$ .*

The converse to Lemma 3 does not hold, in the sense that  $\mathcal{M}(\mathcal{G}_{a \cup c}) \xrightarrow{\text{coll}} c$  and  $\mathcal{M}(\mathcal{G}) \xrightarrow{\text{coll}} c$  do not together imply that  $\mathcal{M}(\mathcal{G}) \xrightarrow{\text{coll}} a \cup c$ . For an illustration consider Figure 1a, which shows the independence graph of the loglinear model with generators  $AB, AC, BD, CD$ . Let  $a = \{B, C\}$  and  $c = \{D\}$ . By Theorem 1,  $\mathcal{M}(\mathcal{G}) \xrightarrow{\text{coll}} c$  and  $\mathcal{M}(\mathcal{G}_{a \cup c}) \xrightarrow{\text{coll}} c$  but  $\mathcal{M}(\mathcal{G}) \not\xrightarrow{\text{coll}} a \cup c$  since  $\{A\}$  is not simplicial in  $\mathcal{G}$ . This confirms that the converse to Lemma 3 cannot hold.

Figure 1 about here.

We now consider CG–regression models.

**Lemma 4** *If  $\mathcal{M}(\mathcal{G}) \stackrel{\text{coll}}{\hookrightarrow} a \cup c$  then  $\mathcal{M}(\mathcal{G})^c \stackrel{\text{coll}}{\hookrightarrow} a \cup c$ .*

**Proof:** For any  $f_{a \cup b|c} \in \mathcal{M}(\mathcal{G})^c$  we know that there exists a  $f' \in \mathcal{M}(\mathcal{G})$  such that  $f'_{a \cup b|c} = f_{a \cup b|c}$ . Since  $f'_{a \cup c} \in \mathcal{M}(\mathcal{G}_{a \cup c})$ , it follows that  $f_{a|c} = f'_{a|c} \in \mathcal{M}(\mathcal{G}_{a \cup c})^c$ , as required.

The converse is not true. For an illustration consider again Figure 1. In Figure 1b is shown a graph  $\mathcal{G}$  for which  $\mathcal{M}(\mathcal{G}) \not\stackrel{\text{coll}}{\hookrightarrow} \{I, X\}$  by Theorem 1, since  $\{J\}$  is not strong. However, if we set  $c = \{X\}$ , we know by the conditional independence restrictions that  $f = f_{I|c} f_{J|c} f_c$  for any  $f \in \mathcal{M}(\mathcal{G})$  so that the CG–regression model  $\mathcal{M}(\mathcal{G})^c$  is obviously equivalent to two independent polytomous logistic regressions, that of  $I$  on  $X$  and of  $J$  on  $X$ . Hence  $\mathcal{M}(\mathcal{G})^c \stackrel{\text{coll}}{\hookrightarrow} \{I, X\}$ , so the converse to Lemma 4 cannot hold. However, with an extra condition we can obtain the reverse implication:

**Lemma 5** *If  $\mathcal{M}(\mathcal{G})^c \stackrel{\text{coll}}{\hookrightarrow} a \cup c$  and  $\mathcal{M}(\mathcal{G}) \stackrel{\text{coll}}{\hookrightarrow} c$  then  $\mathcal{M}(\mathcal{G}) \stackrel{\text{coll}}{\hookrightarrow} a \cup c$ .*

**Proof:** For any  $f \in \mathcal{M}(\mathcal{G})$  we have that  $f_{a|c} \in \mathcal{M}(\mathcal{G}_{a \cup c})^c$  and that  $f_c \in \mathcal{M}(\mathcal{G}_c)$ . From Lemma 2,  $\mathcal{M}(\mathcal{G}_{a \cup c}) \stackrel{\text{coll}}{\hookrightarrow} c$ , and so  $\mathcal{M}(\mathcal{G}_{a \cup c}) = \mathcal{M}(\mathcal{G}_c) \times \mathcal{M}(\mathcal{G}_{a \cup c})^c$ , and thus  $f_c f_{a|c} = f_{a \cup c} \in \mathcal{M}(\mathcal{G}_{a \cup c})$  as required.

Note that Lemma 5 is not applicable to the graph shown in Figure 1b, since here  $\mathcal{M}(\mathcal{G}) \not\stackrel{\text{coll}}{\hookrightarrow} \{X\}$  because  $\{I, J\}$  is not strong in  $\mathcal{G}$ .

To obtain conditions that are both necessary and sufficient for  $\mathcal{M}(\mathcal{G})^c \stackrel{\text{coll}}{\hookrightarrow} a \cup c$  we need to examine more specific aspects of the conditional models: firstly, when distributional properties are preserved under marginalisation and then when Markov properties are preserved.

To characterise when the distributions in  $\mathcal{M}(\mathcal{G})^c_{a \cup c}$  are CG–regressions we must consider in more detail the parametric form of CG–regressions described above

in (2)-(4). Firstly, it is straightforward to see that marginalising over the continuous components  $Y_b$  only affects the mean vector (3) and the covariance matrix (4). The former is reduced to the components  $\xi(i_{a \cup b} | i_c, y_c)_{a \cap \Gamma}$  corresponding to  $Y_a$  and the latter is reduced to the appropriate submatrix  $\Sigma(i_{a \cup b} | i_c, y_c)_{a \cap \Gamma}$ . The structure of the parameters obviously remains the same,

$$\xi(i_{a \cup b} | i_c, y_c)_{a \cap \Gamma} = c(i_{a \cup b} | i_c)_{a \cap \Gamma} + C(i_{a \cup b} | i_c)_{a \cap \Gamma} y_c \quad (6)$$

$$\Sigma(i_{a \cup b} | i_c, y_c)_{a \cap \Gamma} = D(i_{a \cup b} | i_c)_{a \cap \Gamma}, \quad (7)$$

where  $c(i_{a \cup b} | i_c)_{a \cap \Gamma}$  is a  $|a \cap \Gamma|$ -vector,  $C(i_{a \cup b} | i_c)_{a \cap \Gamma}$  a  $|a \cap \Gamma| \times |c \cap \Gamma|$ -matrix, and  $D(i_{a \cup b} | i_c)_{a \cap \Gamma}$  a  $|a \cap \Gamma| \times |a \cap \Gamma|$ -matrix. This implies that marginalising a CG-regression over continuous response variables preserves the distributional form.

In the next lemma we give necessary and sufficient conditions for this when additionally marginalising over discrete variables.

**Lemma 6** *Let  $a, b, c$  be a partition of  $\mathcal{V}$  and suppose that  $X_{a \cup b} = (I_a, I_b, Y_a, Y_b)$  follows a CG-regression given  $X_c = (I_c, Y_c)$ . Then  $X_a = (I_a, Y_a)$  follows a CG-regression given  $X_c$  if and only if there exists a partition  $b = b_1 \cup b_2$ , where either  $b_1$  or  $b_2$  may be empty, such that (i)  $X_{a \cup b_1} \perp\!\!\!\perp X_{b_2} \mid X_c$ , and (ii)  $Y_{a \cup c} \perp\!\!\!\perp I_{b_1} \mid I_{a \cup c}$ .*

**Proof:** For sufficiency, write the conditional density of  $X_{a \cup b}$  given  $X_c$  as  $f_{a \cup b | c}$ . From (i) we have that  $f_{a \cup b | c} = f_{a \cup b_1 | c} f_{b_2 | c}$  so clearly marginalisation over  $b_2$  does not affect the parametric form of  $f_{a \cup b_1 | c}$ . From Lauritzen (1996, section 6.5.1) there exists a joint CG-distribution of  $X = (X_a, X_{b_1}, X_c)$ , say  $f'$ , such that  $f_{a \cup b_1 | c}$  derives from  $f'$  by conditioning. By Lemma 4.1 of Frydenberg (1990b), if (ii) holds,  $f'_{a \cup c}$  is CG and so  $f_{a | c} = f'_{a | c}$  is a CG-regression.

Now let  $b_2$  be the maximal subset of  $b$  such that condition (i) holds and  $b_1 = b \setminus b_2$ . We observe that if  $b_1 \neq \emptyset$ , marginalisation over  $b_1$  may, in general, destroy the required parametric form of  $f_{a | c}$ , or, more specifically, the existence of a parametrisation via  $(\tilde{u}, \tilde{v}, \tilde{W}, \tilde{c}, \tilde{C}, \tilde{D})$  analogous to (2), (3), and (4). Next,

we show that condition (ii) is necessary for this not to happen.

Firstly, for the continuous response variables  $Y_a$  to be CG-distributed after marginalisation over  $b_1$ , we require that  $Y_a \perp\!\!\!\perp I_{b_1} \mid I_{a \cup c}$ , using Frydenberg (1990b, Lemma 4.1). This condition ensures that marginalisation does not yield a mixture of normal distributions for the continuous components. Thus, the mean vector and covariance matrix are given as in (6) and (7) with  $\tilde{c}$ ,  $\tilde{C}$ , and  $\tilde{D}$  not depending on  $i_{b_1}$  (and not on  $i_{b_2}$  because of condition (i)).

Secondly, we show that if the first moment characteristic of the marginal CG-regression of  $X_a$  on  $X_c$  ignoring  $X_b$  can be written analogously to (2) as

$$\log p(i_a | i_c, y_c) = \tilde{u}(i_a | i_c) + \tilde{v}(i_a | i_c)^\top y_c - y_c^\top \tilde{W}(i_a | i_c) y_c - \log \tilde{\kappa}(i_c, y_c)$$

then this implies that  $Y_c \perp\!\!\!\perp I_{b_1} \mid I_{a \cup c}$ . To see this, observe that by (2)

$$\begin{aligned} \log p(i_a | i_c, y_c) &= \log \sum_{i_{b_1}} p(i_{a \cup b_1} | i_c, y_c) \\ &= \log \left[ \sum_{i_{b_1}} \exp \left\{ u(i_{a \cup b_1} | i_c) + v(i_{a \cup b_1} | i_c)^\top y_c - y_c^\top W(i_{a \cup b_1} | i_c) y_c - \log \kappa(i_c, y_c) \right\} \right] \\ &= \log \left[ \sum_{i_{b_1}} \frac{\exp \{ u(i_{a \cup b_1} | i_c) \} \exp \{ v(i_{a \cup b_1} | i_c)^\top y_c \}}{\exp \{ y_c^\top W(i_{a \cup b_1} | i_c) y_c \} \kappa(i_c, y_c)} \right]. \end{aligned}$$

The required structure regarding the linear effects  $v$  and the quadratic effects  $W$  of the continuous  $Y_c$  can therefore only be preserved if  $v(i_{a \cup b_1} | i_c)$  and  $W(i_{a \cup b_1} | i_c)$  are independent of  $i_{b_1}$  so that

$$\begin{aligned} \log p(i_a | i_c, y_c) &= \\ &= \log \left[ \frac{\exp \{ v(i_a | i_c)^\top y_c \}}{\exp \{ y_c^\top W(i_a | i_c) y_c \} \kappa(i_c, y_c)} \sum_{i_{b_1}} \exp \{ u(i_{a \cup b_1} | i_c) \} \right] \\ &= v(i_a | i_c)^\top y_c - y_c^\top W(i_a | i_c) y_c - \log \kappa(i_c, y_c) + \log \left[ \sum_{i_{b_1}} \exp \{ u(i_{a \cup b_1} | i_c) \} \right]. \end{aligned}$$

That this has the required structure can be seen by setting  $\tilde{v} = v$ ,  $\tilde{W} = W$ ,  $\tilde{\kappa} = \kappa$ , and

$$\tilde{u}(i_a | i_c) = \log \left[ \sum_{i_{b_1}} \exp \{ u(i_{a \cup b_1} | i_c) \} \right].$$

If  $v(i_{a \cup b_1} | i_c)$  and  $W(i_{a \cup b_1} | i_c)$  are independent of  $i_{b_1}$  then  $I_{b_1} \perp\!\!\!\perp Y_c \mid I_{a \cup c}$  since

$$\frac{p(i_{a \cup b_1} | i_c, y_c)}{p(i_a | i_c, y_c)} = \frac{\exp\{u(i_{a \cup b_1} | i_c)\}}{\sum_{i_{b_1}} \exp\{u(i_{a \cup b_1} | i_c)\}}$$

is independent of  $Y_c$ . This completes the proof of Lemma 6.

From the above proof it can be seen that marginalising over a discrete variable does not only cause continuous variables in the boundary to have a mixture of normal distributions, as in the unconditional case, but also changes the type of regression for discrete variables in the boundary to a non-logistic type regression.

In terms of models rather than distributions, we can reformulate the lemma as follows.

**Lemma 7** *All distributions in  $\mathcal{M}(\mathcal{G})_{a \cup c}^c$  are CG-regressions given  $X_c = x_c$  if and only if for each connected component  $b_i$  of  $b$  in  $\mathcal{G}$ , either  $\text{bd}(b_i) \subseteq c$  or  $b_i$  is strong in  $\mathcal{G}^c$ .*

**Proof:** This follows from the observation that  $I_b \perp\!\!\!\perp Y_{a \cup c} \mid I_{a \cup c}$  for all  $\mathcal{G}^c$ -Markovian distributions if and only if  $b$  is a strong collection in  $\mathcal{G}^c$  (see the proof of Theorem 4.3 in Frydenberg, 1990b).

We now turn to the preservation of Markov properties under marginalisation.

**Lemma 8** *Let  $a, b, c$  be a partition of  $\mathcal{V}$ . Then  $\mathcal{M}(\mathcal{G})_{a \cup c}^c$  is  $\mathcal{G}_{a \cup c}^c$ -Markovian if and only if  $b$  is a simplicial collection in  $\mathcal{G}^c$ .*

**Proof:** The condition is necessary and sufficient for  $\mathcal{M}(\mathcal{G}^c)_{a \cup c}$  to be  $\mathcal{G}_{a \cup c}^c$ -Markovian. Since this includes no restrictions on the distribution of  $X_c$  it is also equivalent to  $\mathcal{M}(\mathcal{G}^c)_{a \cup c}^c$  being  $\mathcal{G}_{a \cup c}^c$ -Markovian in the sense of (5). By our earlier reasoning we know that  $\mathcal{M}(\mathcal{G})^c = \mathcal{M}(\mathcal{G}^c)^c$ . Thus,  $\mathcal{M}(\mathcal{G})_{a \cup c}^c$  is  $\mathcal{G}_{a \cup c}^c$ -Markovian if and only if  $b$  is a simplicial collection in  $\mathcal{G}^c$ .

The following theorem brings the foregoing lemmas together and is the main result of the paper.

**Theorem 2** *Let  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  be an undirected graph and  $a, b, c$  a partition of  $\mathcal{V}$ .  $\mathcal{M}(\mathcal{G})^c \xrightarrow{\text{coll}} a \cup c$  if and only if, for each connected component  $b_i$  of  $b$ , either (i)  $\text{bd}(b_i) \subseteq c$ , or (ii)  $b_i$  is strong and simplicial in  $\mathcal{G}^c$ .*

**Proof:** This follows directly from Lemmas 7 and 8. Note that (i) implies that  $b_i$  is simplicial in  $\mathcal{G}^c$ , so (i) and (ii) together imply that  $b$  is a simplicial collection in  $\mathcal{G}^c$  and Lemma 8 applies. Only when (i) is violated do we need the additional strongness as formulated in Lemma 7.

Examples of the use of Theorem 2 will be given in the next section.

## 4 Implications for likelihood inference

As mentioned earlier, collapsibility in the unconditional case,  $\mathcal{M}(\mathcal{G}) \xrightarrow{\text{coll}} a$ , is equivalent to  $\mathcal{M}(\mathcal{G}) = \mathcal{M}(\mathcal{G}_a) \times \mathcal{M}(\mathcal{G})^a$ , i.e.  $X_a$  is a so-called cut in  $\mathcal{M}(\mathcal{G})$  (Barndorff-Nielsen, 1978, p. 50). In general, we say that a statistic  $T$  is a cut in a model  $\mathcal{M}$  if it factorises into the product of the marginal model  $\mathcal{M}_T$  and the conditional model  $\mathcal{M}^T$ , where  $\mathcal{M}_T$  and  $\mathcal{M}^T$  are variation independent. Such a property has obvious advantages regarding likelihood inference as it implies, for instance, that  $T$  is sufficient for  $\mathcal{M}_T$ . A similar result is shown in the corollary below for the conditional case. We find that  $\mathcal{M}(\mathcal{G})^c \xrightarrow{\text{coll}} a \cup c$  is equivalent to  $\mathcal{M}(\mathcal{G})^c = \mathcal{M}(\mathcal{G}_{a \cup c})^c \times \mathcal{M}(\mathcal{G})^{a \cup c}$ . More precisely, any density  $f_{a \cup b | c}(\cdot; \theta) \in \mathcal{M}(\mathcal{G})^c$  with regression parameter  $\theta$  can be factorised as

$$f_{a \cup b | c}(x_{a \cup b} | x_c; \theta) = f_{a | c}(x_a | x_c; \phi) f_{b | a \cup c}(x_b | x_{a \cup c}; \gamma),$$

where  $\phi$  and  $\gamma$  are variation independent and represent the regression parameters for the regression of  $X_a$  onto  $X_c$  and  $X_b$  onto  $X_{a \cup c}$ , respectively. We may therefore regard the statistic  $X_a$  as a cut in the conditional model  $\mathcal{M}(\mathcal{G})^c$  where  $x_c$  is given. Note that this is also equivalent to  $X_a$  being S-sufficient for  $\phi$  and S-ancillary for  $\gamma$  given  $x_c$  (Barndorff-Nielsen, 1978, p. 50).

In order to also address the implications for maximum likelihood estimation, we require some additional notation. Let  $\hat{f}$  denote the maximum likelihood estimator in a model, in particular  $\hat{f}_{a \cup b | c}$  is the maximum likelihood estimator in

$\mathcal{M}(\mathcal{G})^c$ . We need to distinguish on the one hand  $\hat{f}_{a|c}$ , the maximum likelihood estimator derived from marginalising  $\hat{f}_{a \cup b|c}$  based on the full model and the full data, and on the other hand  $\hat{f}_{[a|c]}$ , which denotes the maximum likelihood estimator in the model  $\mathcal{M}(\mathcal{G}_a)^c$  based on the measurements  $(x_{a \cup c}^1, \dots, x_{a \cup c}^n)$  of  $a$  and  $c$ , only. A further result shown below is that the two are equal if and only if  $\mathcal{M}(\mathcal{G})^c \xrightarrow{\text{coll}} a \cup c$ .

We now summarize the different equivalent characterisations of collapsibility in CG–regressions.

**Corollary 1** *Let  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  be an undirected graph and  $a, b, c$  a partition of  $\mathcal{V}$ . Then the following are equivalent:*

- (i)  $\mathcal{M}(\mathcal{G})^c \xrightarrow{\text{coll}} a \cup c$ .
- (ii) *For each connected component  $b_i$  of  $b$ , either  $\text{bd}(b_i) \subseteq c$ , or  $b_i$  is strong and simplicial in  $\mathcal{G}^c$ .*
- (iii)  $X_A$  is a cut in  $\mathcal{M}(\mathcal{G})^c$ .
- (iv) *If the maximum likelihood estimator  $\hat{f}_{a \cup b|c}$  in  $\mathcal{M}(\mathcal{G})^c$  exists then  $\hat{f}_{a|c} = \hat{f}_{[a|c]}$ .*

**Proof:** The equivalence of (i) and (ii) is asserted by Theorem 2.

To see that (i) implies (iii), let  $s = \text{bd}(b)$ . Then  $\{(a \cup c) \setminus s; s; b\}$  is a decomposition (in the sense of Leimer, 1989) of  $\mathcal{G}^c$ . This implies that  $X_{a \cup c}$  is a cut in  $\mathcal{M}(\mathcal{G}^c)$  (Lauritzen, 1996, Proposition 6.15) so that  $f \in \mathcal{M}(\mathcal{G}^c)$  factorises into  $f_{a \cup c}$  and  $f_{b|s}$ , where  $f_{a \cup c}$  and  $f_{b|s}$  are variation independent. This is preserved if we further condition on  $x_c$  which affects only  $f_{a \cup c}$ , so that  $f_{a \cup b|c} = f_{a|c} f_{b|s}$ , where  $f_{a|c}$  and  $f_{b|s}$  are again variation independent.

It can be shown in analogy to Frydenberg (1990b, Theorem 5.4) that (iii) implies (i), (i) and (iii) together imply (iv) and that (iv) implies (i), which completes the proof.

The equivalence between collapsibility and certain subsets of variables forming a cut suggests the distinction between two types of CG–regression mod-

els: those in which the set of response variables constitutes a strong collection in  $\mathcal{G}$ , and those in which it does not. We can call the former *strong* and the latter *weak*. When  $\mathcal{M}(\mathcal{G})^c$  is strong,  $\mathcal{M}(\mathcal{G}^c) \xrightarrow{\text{coll}} c$  by Theorem 1, and so  $\mathcal{M}(\mathcal{G}^c) = \mathcal{M}(\mathcal{G}_c^c) \times \mathcal{M}(\mathcal{G})^c$  as discussed above. It follows that parameter estimation for a strong CG–regression model can be performed relatively easily, by fitting the joint undirected model  $\mathcal{M}(\mathcal{G}^c)$  and then deriving the conditional from the joint distribution (Edwards, 2000, p. 87).

Weak CG–regression models cannot be fitted in this way, however. For these the TM algorithm (Edwards and Lauritzen, 2001; Sundberg, 2002) may be used. Computationally, this is very similar to the EM algorithm, involving iteration between two steps: a T–step in which conditional expectations are calculated, and an M–step in which a joint undirected model is fit. However, with property (iii) of the above corollary,  $\mathcal{M}(\mathcal{G})^c \xrightarrow{\text{coll}} a \cup c$  is equivalent to the decomposition of  $\mathcal{M}(\mathcal{G})^c$  into two CG–regression models,  $\mathcal{M}(\mathcal{G})^c = \mathcal{M}(\mathcal{G}_{a \cup c})^c \times \mathcal{M}(\mathcal{G})^{a \cup c}$ . Note that the former may be weak but that the latter, whenever condition (ii) of Theorem 2 holds for each connected component of  $b$ , will be strong. Thus collapsibility allows a decomposition into two lower-dimensional CG–regression models, and often the second of these may be easily estimated.

Various methods have been proposed to speed up the EM algorithm, in particular, the computation of conditional expectations, in the context of mixed graphical models (Lauritzen, 1995; Didelez and Pigeot, 1998; Geng et al., 2000); these may also be applied to speed up the TM algorithm. Exploiting collapsibility may allow gains in efficiency whichever method is used.

**Examples:**

The following examples illustrate the use of Theorem 2 and Corollary 1.

*Figure 2 about here*

All the CG–regressions shown in Figure 2 are weak as the explanatory variable

is continuous and at least one response is discrete.

The CG–regression model shown as Figure 2a is collapsible onto  $\{I, X\}$  because condition (ii) of Theorem 2 applies:  $Y$  is continuous and its boundary  $\{I, X\}$  is complete, so it is strong and simplicial. Hence the model may be decomposed into a binary or polytomous logistic regression of  $I$  on  $X$  and a normal linear regression of  $Y$  on  $I$  and  $X$ . This is a considerable conceptual simplification. It allows the use of ordinary univariate regression methods to examine the dependence of  $Y$  on  $I$  and  $X$ . For example, when the joint model is homogeneous we can use F–tests to examine whether the edges  $[IY]$  or  $[XY]$  may be removed. The model is not collapsible onto  $\{X, Y\}$  because  $\text{bd}(\{I\}) \not\subseteq \Delta$ , so that  $I$  is not strong. Hence it is not equivalent to a linear regression of  $Y$  on  $X$ , followed by a logistic regression of  $I$  on  $X$  and  $Y$ .

The model shown in Figure 2b is collapsible onto  $\{I, X\}$  since  $\text{bd}(\{J\})$  is discrete. Ordinary contingency table methods can be used to examine the dependence of  $J$  on  $I$ . Furthermore, the decomposition may be exploited in estimation to improve efficiency. Let us illustrate this with a numeric example. We generate  $n = 200$  observations from a homogenous CG–distribution for  $(X, I, J)$ ,  $I, J \in \{1, 2\}$ , with cell probabilities  $p(1, 1) = 0.3$ ,  $p(1, 2) = 0.1$ ,  $p(2, 1) = 0.2$ ,  $p(2, 2) = 0.4$ , conditional means  $\xi_X(1, 1) = \xi_X(1, 2) = 3$ ,  $\xi_X(2, 1) = \xi_X(2, 2) = 6$ , and variance  $\sigma_X^2 = 1$  implying that  $X \perp\!\!\!\perp J \mid I$  as in Figure 2b. The summary statistics are given in Table 1.

*Table 1 about here*

The above CG–model leads to a homogenous CG–regression of  $(I, J)$  onto  $X$  with (log) cell probabilities as follows (cf. Lauritzen, 1996, p. 165)

$$\log p(i, j) = u(i, j) + v(i, j)x - w(i, j)x^2 - \log \kappa(x), \quad (8)$$

where  $u(i, j) = \log p(i, j) - \log(2\pi)/2 - \xi_X(i)^2/2$ ,  $v(i, j) = \xi_X(i)$ ,  $w(i, j) = 0.5$  and  $\kappa(x)$  the normalising constant. As  $w(i, j)$  does not depend on  $(i, j)$ , due to homogeneity, it can be subsumed into  $\log \kappa(x)$ .

Fitting a homogenous CG–regression for the model in Figure 2b to the simulated data with MIM (Edwards, 2000) requires 8 iterations of the TM algorithm and yields the estimates given in Table 2 (the model is reparameterised using  $(i, j) = (1, 1)$  as reference category).

*Table 2 about here*

Fitting a homogenous CG–regression of only  $I$  on  $X$  takes again 8 iterations but the computations are faster. Complementing this by a saturated loglinear for  $(I, J)$  results in the maximum likelihood estimators given in Table 3. Note that the conditional probabilities  $P(J = j|I = i)$  can explicitly be estimated from the data in Table 1. It can easily be seen that the estimators from the decomposed model are the same as those obtained by marginalising the results given in Table 2 and vice versa, corroborating part (iv) of Corollary 1.

*Table 3 about here*

Note that the model in Figure 2b does not imply that  $X \perp\!\!\!\perp J$  marginally. Consequently, if we decompose the CG–regression wrongly into a regression of  $J$  on  $X$ , which would be the independence model, followed by a regression of  $I$  on  $(J, X)$  we would get different results not agreeing with those in Table 2 and 3.

The CG–regression model in Figure 2c is not collapsible onto  $\{I, X\}$  as the boundary of  $J$  is not discrete. It follows that it is *not* equivalent to a decomposition into the logistic regression of  $I$  on  $X$  and the logistic regression of  $J$  on  $I$  and  $X$ . Nor is it equivalent to the logistic regression of  $J$  on  $X$  followed by the logistic regression of  $I$  on  $J$  and  $X$ . All three models are distinct. For illustration, we generate again  $n = 200$  observations from a homogenous CG–distribution for  $(X, I, J)$ ,  $I, J \in \{1, 2\}$ , with cell probabilities as before but conditional means  $\xi_X(1, 1) = 2$ ,  $\xi_X(1, 2) = \xi_X(2, 1) = 4$ ,  $\xi_X(2, 2) = 8$ , and variance  $\sigma_X^2 = 1$ . The summary statistics are given in Table 4.

*Table 4 about here*

The homogeneous CG–regression of  $(I, J)$  onto  $X$  has the same form as model (8) except that  $v(i, j) = \xi_X(i, j)$  depends on both discrete variables. Fitting this model with MIM yields the maximum likelihood estimators shown in Table 5. From these we calculate the marginal probability by summing over  $j = 1, 2$

$$\hat{P}(I = 2|X = x) = (\exp\{-6.478 + 2.154x\} + \exp\{-23.847 + 5.303x\})\kappa(x)^{-1},$$

which, as expected, is not a CG–regression as it cannot be represented in the form of model (2).

*Table 5 about here*

However, if the model is wrongly decomposed into the CG–regression of  $I$  on  $X$  we find that a conditional probability

$$\log \hat{P}(I = 2|X = x) = -4.268 + 1.152x - \log \tilde{\kappa}(x)$$

is estimated. The two models yield different conditional probabilities as shown in Table 6, illustrating part (iv) of Corollary 1. Note that the correct model is more accurate for extreme values of  $x$ .

*Table 6 about here*

## 5 Decomposable CG–regressions

In general, it is clearly important to know whether a CG–regression model  $\mathcal{M}(\mathcal{G})^c$  can be decomposed into a sequence of univariate CG–regressions. In this case there exists a (not necessarily unique) ordering  $(\alpha_1, \dots, \alpha_K)$  of the elements in  $a = \mathcal{V} \setminus c$  such that

$$f_{a|c} = \prod_{j=1}^K f_{\alpha_j|\alpha_1, \dots, \alpha_{j-1}, c}, \quad (9)$$

where  $f_{\alpha_j|\alpha_1, \dots, \alpha_{j-1}, c} \in \mathcal{M}(\mathcal{G}_{\{\alpha_1, \dots, \alpha_j, c\}})^{\{\alpha_1, \dots, \alpha_{j-1}, c\}}$  for  $j = 1 \dots k$ . Note that any multivariate distribution can be written as the product of univariate regressions. The point, here, is that these univariate regressions are again in the class

of CG–regressions. In analogy with the unconditional case, we call models with property (9) *strongly decomposable*, and we call an ordering satisfying (9) an SD–ordering. For CG–distribution models characterised by undirected graphs, necessary and sufficient conditions for strong decomposability are well–known (Lauritzen, 1996, Proposition 2.17, p. 18). These are that the graph is *triangulated* and contains no *forbidden paths*, which are defined as paths between two non–adjacent discrete vertices passing through only continuous vertices. For CG–regression models these conditions have to be modified, as we see shortly. We first need a lemma:

**Lemma 9** *If  $\mathcal{M}(\mathcal{G})^c$  is strongly decomposable, then in any connected component of  $a = \mathcal{V} \setminus c$  there can be at most one discrete variable adjacent to a variable in  $c \cap \Gamma$ . Furthermore, in any SD–ordering, this variable will be prior to the other variables in the connected component.*

**Proof:** Suppose for simplicity that  $a$  is connected, and that  $\alpha_i$  and  $\alpha_j$  are two such discrete variables in  $a$  adjacent to  $c \cap \Gamma$ . Let  $(\alpha_1, \dots, \alpha_i, \dots, \alpha_j, \dots, \alpha_K)$  be an SD–ordering of  $a$ , so that  $\mathcal{M}(\mathcal{G})^c$  is collapsible over  $\alpha_K$ , the marginal model is collapsible over  $\alpha_{K-1}$  and so forth. There exists a path between  $\alpha_i$  and  $\alpha_j$  in  $\mathcal{G}_a$  because they are in the same connected component. Since  $\alpha_K$  must be simplicial in  $\mathcal{G}_a$ , there must also exist a path between  $\alpha_i$  and  $\alpha_j$  in  $\mathcal{G}_{a \setminus \{\alpha_K\}}$ , and, continuing in this fashion, also in  $\mathcal{G}_{\{\alpha_1, \dots, \alpha_i, \dots, \alpha_j\}}$ . So  $\text{bd}(\{\alpha_j\}) \not\subseteq c$ , and the strongness property of Theorem 2(ii) is violated. Thus there is only one such variable, say  $\alpha_i$ . The same reasoning applied to  $\alpha_1$  and  $\alpha_i$  shows that such a variable, if it exists, must be the first in the ordering.

Before stating the final theorem we need some more notation. If  $\mathcal{G}^c$  is the graph where  $\mathcal{G}_c$  has been made complete then let  $\tilde{\mathcal{G}}^c$  be the graph where in addition all variables in  $c$  are changed to be discrete. Combining the unconditional and the regression case then yields the following result.

**Theorem 3** *The CG–regression model  $\mathcal{M}(\mathcal{G})^c$  is strongly decomposable if and only if*

- (i) *the graph  $\tilde{\mathcal{G}}^c$  is triangulated and contains no forbidden paths, and*

(ii) for all connected components  $a_j$ ,  $j = 1, \dots, J$ , of  $\mathcal{V} \setminus c$  there is at most one discrete variable in  $a_j$  which is adjacent to  $c \cap \Gamma$  in  $\mathcal{G}^c$ .

**Proof:** For sufficiency, we show that if (i) and (ii) are satisfied then we can find an ordering  $(\alpha_1, \dots, \alpha_K)$  such that the factorisation (9) is valid.

Let  $a = \mathcal{V} \setminus c$ . Consider the situation  $\Gamma \cap a = \emptyset$ , so that there are no continuous response variables. If  $\mathcal{G}^c$  is complete then either the set  $a$  consists only of one vertex or  $c \cap \Gamma = \emptyset$ , so that there are no continuous vertices among the explanatory variables, because otherwise condition (ii) cannot be satisfied. In both cases strong decomposability is trivial and the ordering, in the second case, can be chosen arbitrarily. Assume therefore that  $\mathcal{G}^c$  is not complete. Since  $\mathcal{G}^c$  is triangulated there exist at least two non-adjacent vertices that are simplicial (Lauritzen, 1996, Lemma 2.9, p.13). One of these must be in the set  $a$  as all vertices in  $c$  are adjacent in  $\mathcal{G}^c$ . We can choose this vertex as  $\alpha_K$  (if this is the vertex adjacent to  $c \cap \Gamma$  then, again, for the same reason as before, there cannot be any other vertices in the set  $a$ ). The subgraph  $\mathcal{G}_{\mathcal{V} \setminus \{\alpha_K\}}^c$  is again triangulated (Lauritzen, 1996, Corollary 2.8, p.12) and we can apply the same reasoning to find  $\alpha_{K-1}$ . This procedure can be carried forward until finding  $\alpha_1$  which will be the one vertex possibly adjacent to  $c \cap \Gamma$  if this is non-empty.

Consider now the situation  $\Gamma \cap a \neq \emptyset$ , where there are continuous response variables. Since  $\tilde{\mathcal{G}}^c$  is triangulated and contains no forbidden paths there is at least one strong and simplicial continuous vertex in the set  $a$  (Lauritzen, 1996, Corollary 2.10, p.14). This can be chosen to be  $\alpha_K$ . The subgraph  $\tilde{\mathcal{G}}_{\mathcal{V} \setminus \{\alpha_K\}}^c$  is again triangulated and contains no forbidden paths (Lauritzen, 1996, Corollary 2.8, p.12) so that we can proceed as before if any continuous response variables remain, or as above if only discrete response variables remain.

If the set  $a$  of responses consists of more than one connected component then the foregoing reasoning can be applied separately to these components.

For the reverse, we have to show that if  $\mathcal{M}(\mathcal{G})^c$  is strongly decomposable, then (i) and (ii) follow. That (ii) is necessary follows from Lemma 9. We now show that the undirected model  $\mathcal{M}(\tilde{\mathcal{G}}^c)$  is strongly decomposable, which by

Proposition 2.17 of Lauritzen (1996) is equivalent to (i).

Suppose that  $(\alpha_1, \dots, \alpha_K)$  is an SD-ordering of  $a$  satisfying (9). Suppose also that there are  $J$  connected components of  $a$  and that  $b$ , say, is the subset of  $a$  consisting of the  $J$  initial variables of the connected components under the ordering. By Lemma 9, these are the variables for which the strongness requirement of Theorem 2 condition (ii) may not hold. Consider  $\tilde{G}_{c \cup b}^c = \mathcal{H}$ , say. We have that  $\mathcal{H}_c$  is complete,  $c$  is here regarded as discrete, and the vertices in  $b$  are disconnected in  $\mathcal{H}_b$  and both strong and simplicial in  $\mathcal{H}$ . It follows that the undirected model  $\mathcal{M}(\mathcal{H})$  is strongly decomposable and there exists an SD-ordering of  $c \cup b$  in which the vertices in  $b$  come last. Write this as  $(c_1, \dots, c_{|C|}, b_1, \dots, b_J)$ . Then  $(c_1, \dots, c_{|C|}, \alpha_1, \dots, \alpha_K)$  is an SD-ordering of  $\mathcal{V}$  relative to  $\tilde{G}^c$ , implying that the undirected model  $\mathcal{M}(\tilde{G}^c)$  is strongly decomposable as required.

The above theorem provides necessary and sufficient condition for the decomposition (9) which are very similar to the unconditional case. Note that algorithms that check for decomposability are well known (Cowell et al., 1999, p.134) and can easily be adapted to our case.

### Examples:

The models in Figure 2a and 2b are obviously strongly decomposable. In terms of Theorem 3 we can see that Figure 2c is not strongly decomposable because there are two discrete responses adjacent to a continuous covariate. Note that Figure 2b is triangulated and contains no forbidden path so that if  $X$  was not fixed to be a covariate the model could be decomposed.

*Figure 3 about here*

For slightly more complex examples consider the models in Figure 3. The response variables are  $\{J, K, Y\}$  and the explanatory  $\{I, X\}$ . The model in 3a is not strongly decomposable as it is not triangulated: there is a four-cycle  $(X, J, K, I)$ . The model is collapsible over  $\{Y\}$ , but the regression of  $\{J, K\}$  on  $\{I, X\}$  cannot be further decomposed. Note that  $(I, X, J)$  is not a forbidden

path, here, because the explanatory variables are regarded as discrete when checking for forbidden paths. Adding a link between  $I$  and  $J$ , as in Figure 3b, makes the graph triangulated. As the only possible SD-ordering here is  $(J, K, Y)$ , the decomposition into univariate regressions is given as

$$f_{J,K,Y|I,X} = f_{Y|I,K} f_{K|I,J} f_{J|I,X},$$

where  $f_{Y|I,K}$  is an analysis of variance model and  $f_{K|I,J}$  and  $f_{J|I,X}$  are polytomous logistic regressions with discrete or mixed covariates. The model in Figure 3c is again not decomposable because it contains the forbidden path  $(J, Y, K)$  among the responses.

## 6 Discussion

Our finding shed an interesting light on regression models when latent variables might be relevant. If the model for the full set of variables, including the latent ones, is not collapsible onto the observed ones any standard choice of a regression model for the observed variables, i.e. one in the class of CG-regressions, might be ‘wrong’ in the sense that it will not be the model derived from marginalising the full model. However, if the model is collapsible any parameters relating to the latent variables will not be identifiable from the observed data. A first attempt to provide a general class of graphical models that is closed under marginalisation and conditioning and therefore allows for different types of latent variables has been made by Richardson and Sprites (2002). As yet, however, a parameterisation of these models is only possible in the purely Gaussian case.

Finally, we would like to point out an open problem. The graphical CG-regression models comprise as special cases logistic regression, linear regression and analysis of variance models. One common feature is that they include all higher order interactions consistent with their interaction graph. In many applications this is undesirable since it requires, for instance, the estimation of large numbers of parameters. In order to extend the results to regression models with restricted higher order interactions, we should consider the con-

ditional models derived from hierarchical interaction models (Edwards, 1990). However, the conditions in Theorem 2 are not enough to ensure collapsibility in these models since special care must be taken in order not to destroy the specific parametric restrictions when marginalising. At present, no suitable necessary and sufficient conditions for collapsibility in hierarchical CG–regression models are known. However, recall that all the results presented in this paper are equally valid for homogenous CG–regressions, i.e. under the assumption of constant covariance structure of the continuous variables over the discrete ones, as for instance common in many MANOVA type models. If this seems a reasonable assumption in a given data situation the number of parameters to be estimated can be considerably reduced.

## References

- Asmussen, S., & Edwards, D. (1983). Collapsibility and response variables in contingency tables. *Biometrika* **70**, 566-78.
- Barndorff-Nielsen, O.E. (1978). *Information and exponential families in statistical theory*. Wiley, New York.
- Cowell, R.G., Dawid, A.P., Lauritzen, S.L., & Spiegelhalter, D.J. (1999). *Probabilistic networks and expert systems*. Springer, New York.
- Darroch, J.N., Lauritzen, S.L., & Speed, T.P. (1980). Markov fields and log linear models for contingency tables. *Ann. Statist.* **8**, 522-39.
- Dawid, A.P. (1979). Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. Ser. B* **41**, 1-31.
- Dempster, A.P. (1972). Covariance selection. *Biometrics* **28**, 157-75.
- Didelez, V., & Pigeot, I. (1998). Maximum likelihood estimation in graphical models with missing values. *Biometrika* **85**, 960-6.
- Edwards, D. (1990). Hierarchical interaction models (with discussion). *J. Roy. Statist. Soc. Ser. B* **52**, 3-20.

- Edwards, D. (2000). *Introduction to graphical modelling (2nd edition)*. Springer, New York.
- Edwards, D., & Lauritzen, S.L. (2001). The TM algorithm for maximising a conditional likelihood function. *Biometrika* 88, 961-72.
- Frydenberg, M. (1990a). The chain graph Markov property. *Scand. J. Statist.* **17**, 333-53.
- Frydenberg, M. (1990b). Marginalization and collapsibility in graphical interaction models. *Ann. Statist.* **18**, 790-805.
- Geng, Z., Wan, K. & Tao, F. (2000). Mixed graphical models with missing data and the partial imputation EM algorithm. *Scand. J. Statist.* **27**, 433-44.
- Lauritzen, S.L., & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.* **17**, 31-57.
- Lauritzen, S.L. (1995). The EM algorithm for graphical association models with missing data. *Comput. Statist. Data Anal.* **19**, 191-201.
- Lauritzen, S.L. (1996). *Graphical Models*. Oxford University Press, Oxford.
- Lauritzen, S.L. (1989). Lectures on Contingency Tables (3rd edition). Technical Report R-89-29, Institute for Electronic Systems, Aalborg University.
- Leimer, H.-G. (1989). Triangulated graphs with marked vertices. In *Graph theory in memory of G.A. Dirac*, (ed. L.D. Andersen, C. Thomassen, B. Toft, and P.D. Vestergaard), pp. 311-24. Annals of Discrete Mathematics 41. Elsevier, Amsterdam.
- Madigan, D., and Mosurski, K. (1990). An extension of the results of Asmussen and Edwards on collapsibility in contingency tables. *Biometrika* **77**, 315-9. Correction: *Biometrika* **86**, 973-4.
- Richards, T., Spirtes, P. (2002). Ancestral graph Markov models. *Ann. Statist.* **30**, 962-1030.

Sundberg, R. (2002). The convergence rate of the algorithm of Edwards and Lauritzen. *Biometrika* **89**, 478-83.

Wermuth, N., & Lauritzen, S. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion). *J. Roy. Statist. Soc. Ser. B* **52**, 21-72.

Vanessa Didelez, University College London, Gower Street, London WC1E 6BT, UK.

E-mail: [vanessa@stats.ucl.ac.uk](mailto:vanessa@stats.ucl.ac.uk)

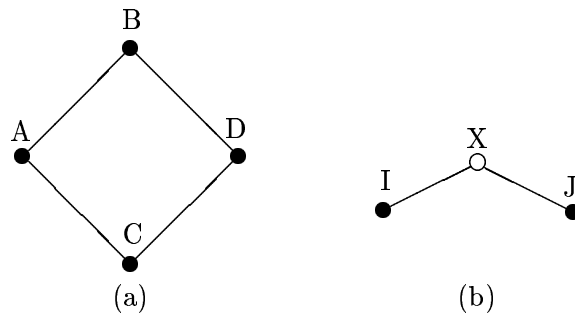


Figure 1: The graph  $\mathcal{G}$  in (a) represents the loglinear model with generators  $AB, AC, BD, CD$ . It implies that  $A \perp\!\!\!\perp D \mid B, C$  and  $B \perp\!\!\!\perp C \mid A, D$ . In (b) is shown a graph for a CG-distribution with one continuous,  $X$ , and two discrete variables,  $I, J$ . The graph implies that  $I \perp\!\!\!\perp J \mid X$ .

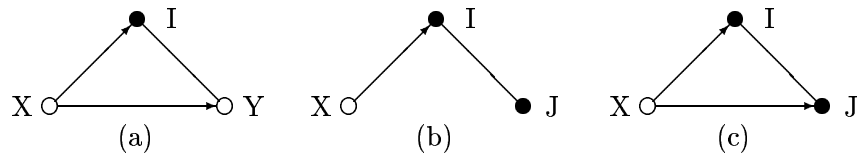


Figure 2: Three CG–regression models. Models (a) and (b) are collapsible onto  $\{I, X\}$  but not (c).

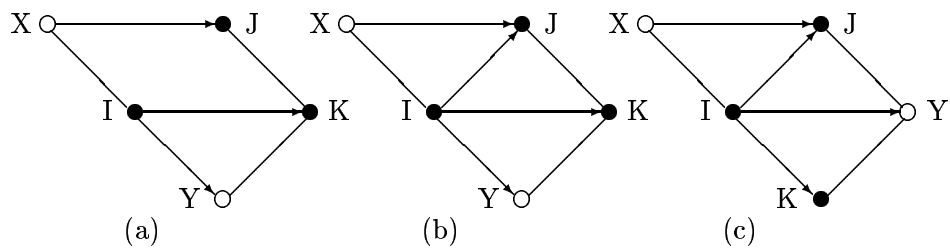


Figure 3: Three CG-regression models. Model (b) is strongly decomposable but neither (a) nor (c) are.

Cell	Mean	Variance	Cell	Mean	Variance
$I, J = 1, 1$ $n = 63$	3.110	0.984	$I, J = 1, 2$ $n = 18$	2.865	1.683
$I, J = 2, 1$ $n = 36$	6.171	0.909	$I, J = 2, 2$ $n = 83$	5.772	1.162

Table 1: Summary statistics for the simulated data on  $(X, I, J)$  with independence structure as in Figure 2b.

$I = i$	$J = j$	intercept	coefficient of $x$
1	2	-1.253	0.000
2	1	-10.872	2.286
2	2	-10.037	2.286

Table 2: Maximum likelihood estimators for the CG–regression as in Figure 2b based on simulated data ( $(I = 1, J = 1)$  is used as reference category).

Estimators for regression $I$ on $X$			Estimators for regression $J$ on $I$		
$I = i$	intercept	coefficient of $x$	$J = j$	$I = i$	cond. probability
2	-9.928	2.286	1	1	0.778
			1	2	0.303

Table 3: Maximum likelihood estimators for decomposed CG-regressions  $I$  onto  $X$  ( $I = 1$  is the reference category) and  $J$  onto  $I$ .

Cell	Mean	Variance	Cell	Mean	Variance
$I, J = 1, 1$ $n = 63$	1.821	0.760	$I, J = 1, 2$ $n = 18$	4.006	1.174
$I, J = 2, 1$ $n = 36$	3.802	1.065	$I, J = 2, 2$ $n = 83$	7.851	1.222

Table 4: Summary statistics for the simulated data on  $(X, I, J)$  with independence structure as in Figure 2c.

$I = i$	$J = j$	intercept	coefficient of $x$
1	2	-7.901	2.341
2	1	-6.478	2.154
2	2	-23.847	5.303

Table 5: Maximum likelihood estimators for the CG-regression as in Figure 2c based on simulated data ( $(I = 1, J = 1)$  is used as reference category)

$X = x$	$\hat{P}_{full}(I = 2 x)$	$\hat{P}_{dec}(I = 2 x)$	$P_{true}(I = 2 x)$
1	0.013024	0.042452	0.011991
2	0.098927	0.123035	0.079469
3	0.410185	0.307464	0.333337
4	0.616603	0.584191	0.587362
5	0.656080	0.816378	0.662765
6	0.882959	0.933640	0.856234

Table 6: Estimated conditional probabilities for given  $X$  values in the (correct) full model,  $\hat{P}_{full}$ , and in the wrongly decomposed model,  $\hat{P}_{dec}$ , as well as the true ones  $P_{true}$ .