Graphical models, regression graphs, and recursive linear regression in a unified way

Marianna Bolla *, Fatma Abdelkhalek †, and Máté Baranyi ‡

Institute of Mathematics, Budapest University of Technology and Economics, 1111. Budapest, Műegyetem rkp. 3, Hungary

February 11, 2019

Dedicated to András Krámli on the occasion of his 75th birthday

This versatile topic goes back to the inventions of Gauss, Markov, and Gibbs, whose ideas are incorporated in graphical models and regression graphs. Later, the geneticist, S. Wright (1923–1934) and the philosopher and computer scientist, J. Pearl (1986–1987) developed the tools, but their notation is too complicated to formulate the mathematical background. Here we mainly follow the up-to-date discussion of statisticians, S. Lauritzen and N. Wermuth, and try to juxtapose the directed–undirected and discrete–continuous cases.

1 Graphical Models in General

First, without specifying the type of distribution, we discuss the directed and undirected models separately. We will show that they have many properties in common, and after possible alterations, can be transformed into each other. If we have a sample from a joint distribution that basically does not have directions between the variables, we can find conditional independences between subsets of them (supporting some kind of Markovity) based on statistical analysis, and build usually an undirected graph on them. In particular, when our underlying distribution is multivariate Gaussian, we build a so-called concentration graph (see Section 3), and if it is discrete on categorical variables with assumed interactions, we build a log-linear model (see Section 2.1). Both models contain decomposable ones as special cases, in which case a so-called perfect numbering of the variables exists. This ordering traces back to the directed case. As a combination of the discrete and continuous, we may have Conditional Gaussian (CG) distributions, and as a combination of the directed and undirected, chain graphs and regression graphs are at our disposal. We want to show that all these

*marib@math.bme.hu
†fatma@math.bme.hu
‡baranyim@math.bme.hu

are strongly related, and only based on the actual data and after a thorough statistical analysis on them, one can determine which model to use.

Summarizing, graphical models provide a framework for describing statistical dependences in (possibly large) collections of random variables. At their core lie various correspondences between the conditional independence properties of a random vector and the structural properties of the graph used to represent its distribution. If there are groups of the variables which are marginally independent, then the joint distribution factorizes trivially. Usually this is not the case, but certain groups of variables can be conditionally independent conditioned on another group. This also causes the joint probability mass function (pmf) or probability density function (pdf) to factorize in terms of certain conditional probabilities or densities. The factors are far not unique, and sometimes they or their negative logarithms are called potentials. Again, the graph here is just a tool for the representation of the more rich structure of the joint distribution.

1.1 Directed Graphical Model: Bayesian Network (BN)

BN's are directed graphical representations of joint distributions. The vertices correspond to random variables (rv's) $X_1, \dots, X_d$, whereas the directed edges to *causal* dependences between them. The rv's are usually discrete, mainly categorical, taking on finitely many values. The point is that even if the rv's are binary, it is time-consuming to learn the underlying distribution from the data as there are 2^d possible joint states in the pmf. However, if we parameterize with the conditional probabilities along the dependences, we can reduce the calculations, provided the underlying distribution $\mathbb{P}$ is Markov compatible with the directed graph assigned to the rv's based on causal relations. But this is sometimes not an open question, as the joint distribution is generated through conditional probability tables (as in [13]), and so, we have it in a factorized form, which fact will turn out to be equivalent to some kind of Markovity.

We treat only *directed acyclic graphs* (DAG's). In case of a DAG G with vertex-set $V = \{1, \dots, d\}$, there are no directed cycles, and therefore, there exists a linear ordering (labeling) of the vertices such that for every directed edge $j \to i$, $i < j$ holds (we can refer to this relation as j is the parent of i). So the youngest vertex has label 1, and the older a vertex, the larger its label is (we can think of labels as ages). We use this, so-called (not necessarily unique) *topological labeling* of the vertices which is also in accord with the labeling of the forthcoming regression graph models, see Section 4.2. To find such a labeling, an algorithm is to be found on page 1146 of [11]. Note, that some authors use the opposite ordering, however, this one fits better in the framework of regression graphs, when it is important that the rows or columns of the involved matrices be indexed in this order.

The *directed factorization* property (DF) of the distribution of the random vector $(X_1, \dots, X_d)$ means that for any state configuration $\mathbf{x} = (x_1, \dots, x_d)$, it factorizes over the DAG G like

$$p(x_1, \dots, x_d) = \prod_{i=1}^{d} p(x_i | x_{i+1}, \dots, x_d) = \prod_{i=1}^{d} p(x_i | \mathbf{x}_{\mathrm{par}(i)}), \qquad \text{(DF)}$$

where $p(x_1, \dots, x_d)$ is the pmf corresponding to the d-tuple of states $(x_1, \dots, x_d)$

in the topological ordering; $\text{par}(i) \subset \{i+1, \dots, d\}$ denotes the set of vertices j such that from them, a directed edge $j \to i$ emanates to i (they are the parents of i), and for any $A \subset V$ we use the notation $\mathbf{x}_A = \{x_i : i \in A\}$ and $\mathbf{X}_A = \{X_i : i \in A\}$.

In fact, (DF) automatically holds if the pmf is constructed based on conditional probabilities (or densities in the continuous case). Vice versa, we can construct a DAG based on a factorized joint density in the following way: we draw a $j \to i$, $i < j$ edge if there is a factor like $p(x_i | \dots x_j \dots)$ that cannot be further reduced.

On the other hand, let (DL) denote the *directed local Markov property* of the distribution of the random vector $\mathbf{X} = (X_1, \dots, X_d)$ as follows. Let

$$\text{ant}(i) = \{i+1, \dots, d\} \setminus \text{par}(i)$$

denote the set of *anteriors* of i (the set of its non-descendants except its parents). Then (DL) means that

$$X_i \perp\!\!\!\perp \mathbf{X}_{\text{ant}(i)} | \mathbf{X}_{\text{par}(i)}, \quad i = 1, \dots, d \tag{DL}$$

holds, i.e., X_i (future) and $\mathbf{X}_{\text{ant}(i)}$ (past) are independent conditioned on $\mathbf{X}_{\text{par}(i)}$ (present). It also means that

$$p(x_i | \mathbf{x}_{\{i+1,\dots,d\}}) = p(x_i | \mathbf{x}_{\text{par}(i)}), \quad i = 1, \dots, d$$

holds for any state configuration. This generalizes the fundamental property of Markov chains (when G is a directed path).

Later on, we need the ancestral set of X_i that consists of the variables X_j, $j \in \{i+1, \dots, d\}$, such that there is a directed path from j to i. This set contains the parents, grandparents, etc. of X_i. For $A \subset V$, let $\text{An}(A)$ denote the *ancestral set* of A, that is the smallest possible vertex-set (including A) containing all vertices from where a directed path emanates to vertices of A.

Theorem 1 of [25] proves that for any DAG G, the set of distributions enjoying property (DF) is the same as those enjoying (DL). Actually, Lauritzen [14] states more. He proves that (DL) is also equivalent to (DG), the global Markov property on directed graphs, but we can define it only in Section 1.2 in the context of undirected graphs and the so-called d-separation. However, we are able to define here the *directed pairwise Markov property* (DP) of the distribution of $(X_1, \dots, X_d)$ that reads as follows:

$$X_i \perp\!\!\!\perp X_j | \mathbf{X}_{\text{par}(i)} \quad \text{for} \quad j \in \text{ant}(i), \quad i = 1, \dots, d. \tag{DP}$$

The (DL)$\Longrightarrow$(DP) implication is trivial, but Lauritzen [14] (page 51) shows in a counterexample that the converse is not always true. Note that Wermuth [31] (page 4) characterizes pairwise dependences too, in addition to the independence statements in (DP). These together are equivalent to (DL). We will summarize these issues in the next sections.

1.2 Undirected Graphical Model and the Markov Random Field (MRF)

Here the vertices also correspond to rv's $X_1, \dots, X_d$, whereas the undirected edges are obtained through conditional independences between them. So MRF's

are undirected graphical models that include the neighbors instead of parents in the conditional independence statements, satisfying some (or all) of the following Markov-type properties.

For an undirected graph G, the *undirected global Markov property* (UG) of a joint distribution with respect to G is defined as follows:

$$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B | \mathbf{X}_S \tag{UG}$$

holds for any vertex cutset S between disjoint vertex-subsets A and B, i.e., removing vertices of S will make A and B disconnected.

The *undirected pairwise Markov property* (UP) of a joint distribution with respect to the undirected graph G is defined as

$$X_i \perp\!\!\!\perp X_j | \mathbf{X}_{V \setminus \{i,j\}}, \quad i \neq j, \tag{UP}$$

while the *undirected local Markov property* (UL) as

$$X_i \perp\!\!\!\perp \mathbf{X}_{V \setminus \mathrm{cl}(i)} | \mathbf{X}_{\mathrm{bd}(i)}, \quad \forall i, \tag{UL}$$

where $\mathrm{bd}(i) = \{j : j \sim i\}$ denotes the set of *neighbors*, in other words, *boundary* (in G) of i, while $\mathrm{cl}(i) = \{i\} \cup \mathrm{bd}(i)$ denotes the *closure* (in G) of i. For the states, Equation (UL) means that

$$p(x_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d) = p(x_i | \mathbf{x}_{\mathrm{bd}(i)}), \quad i = 1, \ldots, d.$$

So the conditional independence relations depend on the neighborhood. This observation is the base of the so-called Gibbs fields.

Now, let (UF) denote the *undirected factorization* property of the underlying multivariate distribution with respect to the undirected graph G. For the states, it means the factorization of the joint pmf or pdf like

$$p(x_1, \ldots, x_d) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \Psi_C(\mathbf{x}_C) \tag{UF}$$

with normalizing constant $Z > 0$ and non-negative *compatibility functions* Ψ_C's assigned to the cliques $C \in \mathcal{C}$ of G. Under *clique* we understand a maximal complete subgraph of G. Note that, in graph theory, they are sometimes called maximal cliques. The compatibility functions are sometimes called *clique potentials*, though this notion is used in the literature in many contexts. In special (to be called decomposable) models, the forthcoming Equation (11) gives an explicit formula for the compatibility functions. The above factorization (UF) is far not unique, and sometimes has a more convenient form if not only the cliques, but other complete subgraphs are also involved (e.g., in hierarchical log-linear models).

In fact, Ψ_C's are defined on all the state configurations within the clique, and depend on the relation of C to the other cliques too. More precisely, $\Psi_C : \mathcal{X}_C \to \mathbb{R}_+$, where $\mathcal{X}_C = \times_{i \in C} \mathcal{X}_i$ and $\mathcal{X}_i$ is the sample space corresponding to X_i, i.e., X_i takes on values in the set $\mathcal{X}_i$. The whole sample space is $\mathcal{X} = \times_{i=1}^d \mathcal{X}_i$.

Lauritzen [14] proves that for a distribution over an undirected graph, the implications

$$\text{(UF)} \Longrightarrow \text{(UG)} \Longrightarrow \text{(UL)} \Longrightarrow \text{(UP)}$$

always hold. However, there is an important theorem, attributed to **Hammersley and Clifford** (see [14, 19]) that states

$$(\text{UP}) \Longrightarrow (\text{UF}),$$

whenever $\mathbb{P}$ is positive and continuous with respect to the product measure which condition always holds in non-degenerate exponential families. So under this condition,

$$(\text{UF}) \Longrightarrow (\text{UG}) \Longrightarrow (\text{UL}) \Longrightarrow (\text{UP}) \Longrightarrow (\text{UF}),$$

therefore, all these properties are equivalent. Consequently,

$$(\text{UG}) \Longleftrightarrow (\text{UL}) \Longleftrightarrow (\text{UP}) \tag{1}$$

also holds. However, for the equivalences in (1), milder conditions than the positivity of $\mathbb{P}$ also suffice. For example, for disjoint vertex-subsets A, B, C:

$$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B | \mathbf{X}_C \quad \text{and} \quad \mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_C | \mathbf{X}_B \Longrightarrow \mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_{B \cup C}.$$

These are the so-called composition and intersection properties that define so-called *graphoid*s and Gaussoids, see [16]. However, such distributions, like the Gaussian, symmetric binary, and so-called MTP_2 distributions, mimic the properties of the Gaussian one, and we do not need the abstract definition of them.

An important consequence of the Hammersley–Clifford theorem is that, in case of positive distributions, (UP) can be used to build the graph that is based on pairwise relations of the variables. Then, with this graph, all the global and local independences will hold.

Gaussian distributions in the continuous, while log-linear distributions in the discrete cases, and the mixture of them all satisfy the positivity constraint, and in the next sections we shall confine ourselves to these distributions. We will show that these distributions have many properties in common, in particular, when the models are decomposable (in the wording of [28], multiplicative), and therefore, the algorithms on a so-called junction tree can be unified in the possession of data. Also, even if the underlying graph is undirected, the decomposable structure gives a (not necessarily unique) so-called *perfect ordering* of the vertices, in which order directed edges can be drawn. This is the base of *path analysis* [33], *regression graph* and *chain graph* models, see [29, 30, 31, 32]. We will go through these topics in Section 4, after discussing log-linear and Gaussian models in Sections 2.1 and 3, as prototypes, in details.

Note that the (UP) $\Longleftrightarrow$ (UG) equivalence in (1) justifies that, in case of positive distributions, all independence statements can be read off the graph, constructed based on pairwise independences of the variables, conditioned on the remaining ones. In Cox and Wermuth [5] these are called *concentration graph*s, but we introduce this notion only in Section 3 for Gaussian rv's. In this context, sometimes *covariance graph*s are considered, where undirected edges correspond to non-zero pairwise correlations. In a concentration graph, when two disjoint vertex-subsets A and B have no path between them, it follows that $\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B$, because $\mathbf{X}_A$ and $\mathbf{X}_B$ are independent conditioned on $\mathbf{X}_\emptyset$ (as there is no separating set between them); this means that they are marginally independent. Otherwise, a covariance graph and concentration graph based on

the same (positive) distribution coincide only if the latter consists exclusively of disjoint cliques. Covariance graphs will be discussed in context of regression graphs (see Section 4.2), but only within chain blocks of rv's on equal standing.

Going back to the directed graphs, observe that condition (DF) resembles that of (UF), since in case of a DAG, condition (DF) can be written as

$$p(x_1, \dots, x_d) = \frac{1}{Z} \prod_{i=1}^{d} \Psi_{\mathrm{cl}(i)}(x_i, x_{\mathrm{par}(i)}) = \frac{1}{Z} \prod_{i=1}^{d} \Psi_{\mathrm{cl}(i)}(\mathbf{x}_{\mathrm{cl}(i)})$$

where $Z = 1$ and $\mathrm{cl}(i) = \{i\} \cup \mathrm{par}(i)$ is considered as the closure of vertex i in the DAG, which also forms a complete subgraph in the skeleton if the DAG G does not contain a *sink* V configuration like $i \to k \leftarrow j$ for $k < i$, $k < j$, when there is no arrow between the distinct vertices i and j. Such a DAG is called *decomposable*. Let us form the $M_i := \mathrm{cl}(i)$ vertex-sets ($i = 1, \dots, d$), the spanning subgraph of which is complete. Delete those that are contained in another one, and keep only the maximal complete subgraphs, i.e., the cliques $C \in \mathcal{C}$ among them. Then the above equation can be transformed into Equation (UF) with Z now not necessarily 1. We will later see that the so obtained cliques also form a so-called junction tree structure in the skeleton (undirected version) of the DAG. This (in other words, decomposable) structure is not necessary understood in Equation (UF) in case of an undirected graph, where the factorization over the cliques does not assume their junction tree structure, so it is weaker than decomposability.

We can make a directed BN undirected: not only disregard the orientation of the edges but also 'moralize' the graph. If G is a DAG , it can be done by connecting two parents (having a common child) whenever they are not connected (married). The so obtained *moral graph* G^m is then used in the MRF setup. To motivate moralization, assume that the underlying distribution is multivariate Gaussian on a DAG G. Moralization is needed when for some triple $i, j > k$ in G, $i \to k$, and $j \to k$ holds, but there is no directed edge between i and j. Then even if X_i and X_j are (marginally) independent, they are not conditionally independent any more, conditioned on X_k. For example, if X_i is the years of former schooling and X_j is the gender, then – though they are independent (men and women can get any education irrespective of gender) – they are conditionally dependent given the income (X_k). In the example of [31], on given level of salary, women had a higher level of education than men. Such conditional dependence induces an edge in Gaussian covariance selection models. The triplet i, k, j is a sink V; for details about these *edge-inducing dependences* see Section 4.1.

Though we may think that an undirected graph gives rise to a richer structure of independence statements through neighborhoods than a directed one (on the same skeleton) using only ancestral dependences, it turns out that the directed and undirected Markov properties are strongly related to each other.

Proposition 1 (Lemma 3.21 of [14]) *If $\mathbb{P}$ has the property (DF) with respect to the directed graph G, then it has the property (UF) with respect to the undirected moral graph G^m of G.*

Certain converse of Proposition 1 and so-called Markov-equivalences of regression graphs will be stated later, when the graph has both directed and undirected

edges, and we have learned the notion of decomposability.

Now, in possession of Proposition 1, we are able to define the *directed global Markov property* (DG) of a distribution with respect to a directed graph G. This means that

$$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B | \mathbf{X}_S \tag{DG}$$

holds for any vertex cutset S between disjoint vertex-subsets A and B in the moral graph of the ancestral set $\mathrm{An}(A \cup B \cup S)$. Note that (DG) is equivalent to the *d-separation* (direction-dependent separation) criterion of Pearl [19]:

$$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B | \mathbf{X}_S \iff S \text{ d-separates } A \text{ from } B, \tag{2}$$

where S d-separates A from B in the DAG G if there is no *active path* in G from A to B given S. A path between A and B given S is active if among its inner nodes, every collision node ($\circ \to \circ \leftarrow \circ$) is in $\mathrm{An}(S)$ and every transmitting node ($\circ \to \circ \to \circ$) is in $V \setminus (A \cup B \cup S)$. This notion is also generalized to regression graphs [29], see Section 4.2. Based on these, the so-called *Bayes-ball* algorithm is constructed to decide whether the above d-separation holds for given disjoint subsets $A, B, S \subset V$, see, e.g., [19]. Later it was shown that the criterion of d-separation cannot be improved: in the case of real sample spaces, a Gaussian and a symmetric binary distribution always exists satisfying (2).

Recall that

$$(\mathrm{DF}) \iff (\mathrm{DG}) \iff (\mathrm{DL}) \implies (\mathrm{DP})$$

as discussed in Section 1.1.

Finally, note that $\mathbb{P}$ is called a Gibbs distribution over the undirected graph G if it can be parameterized by a set of positive functions Ψ_C's over the cliques of G, by physicists called *clique potential*s, such that for its pmf or pdf the condition (UF) holds. By the above Hammersley–Clifford theorem, a Gibbs field and MRF are equivalent with regard to the same G whenever $\mathbb{P}$ is strictly positive. We use the notion Markov Random Field (MRF) only for positive distributions, when all the Markov properties are equivalent to each other and to the factorization property. Originally, Gibbs fields were developed in statistical physics, where the compatibility functions are of the form $\Psi_C = e^{-g_C}$ with g_C an energy function over states $\mathbf{x}_C$ of C. The energy represents the likelihood of the corresponding relationships within the clique, with a higher energy configuration having lower probability and vice versa. The estimation of these potentials through energy functions is related to the theory of the forthcoming log-linear models and Markov Chain Monte Carlo methods, e.g., Gibbs samplers [12, 14]. When the cliques are vertex-pairs (e.g., G is a grid), then we get the classical Ising model.

In [13], the authors give several equivalent potential representations of the probabilities in an MRF, and in the more special class of them (decomposable model), the forthcoming representation of Equations (9) and (11) indeed give a direct factorization of the joint density.

2 Discrete MRF's

2.1 Log-linear models

Let $X_1, \ldots, X_d$ be categorical variables, where X_i takes on values in the finite set $\mathcal{X}_i = \{1, \ldots, r_i\}$, $i = 1, \ldots, d$. The components of the random vector $(X_1, \ldots, X_d)$ are usually not independent, the observations for their joint distribution are collected in a so-called *contingency table*, the frame of which is provided by the sample space (state space) $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$. In fact, $\mathcal{X}$ is a d-dimensional array, the entries of which are d-tuples $\mathbf{x} = (x_1, \ldots, x_d) \in \mathcal{X}$, and they are called *cells*; altogether, there are $\prod_{i=1}^{d} r_i$ cells. Under contingency table we understand this frame together with the cell counts $n(\mathbf{x})$, $\mathbf{x} \in \mathcal{X}$, where the nonnegative integer $n(\mathbf{x})$ is the number of observations for the random vector $\mathbf{X} = (X_1, \ldots, X_d)$ that fall in the cell $\mathbf{x}$ out of the total n observations. In other words, n is the sample size, and of course, $n = \sum_{x \in \mathcal{X}} n(\mathbf{x})$. When n is kept fixed, the joint distribution of the counts, $N(\mathbf{x})'s$ as rv's, is multinomial with parameters $p(\mathbf{x})$, $\mathbf{x} \in \mathcal{X}$:

$$\mathbb{P}(N(\mathbf{x}) = n(\mathbf{x}),\, \mathbf{x} \in \mathcal{X}) = \frac{n!}{\prod_{\mathbf{x} \in \mathcal{X}} n(\mathbf{x})!} \prod_{\mathbf{x} \in \mathcal{X}} p(\mathbf{x})^{n(\mathbf{x})}. \tag{3}$$

In the *saturated model*, the parameters are only constrained by restrictions that are due to the sampling procedure, the multinomial sampling. Under multinomial sampling, as in exponential families, the ML-estimate of the parameters is obtained by equating the count $n(\mathbf{x})$ to the binomial expectation $np(\mathbf{x})$, for all $\mathbf{x} \in \mathcal{X}$, and hence, $\hat{p}(\mathbf{x}) = \frac{n(\mathbf{x})}{n}$, $\quad \mathbf{x} \in \mathcal{X}$.

Now, with some restrictions on the marginal distributions, we shall define more special models. We need the following definitions. The marginal of the contingency table corresponding to a given subset of the variables $\mathbf{X}_A = \{X_i : i \in A\}$, with $A \subset V = \{1, \ldots, d\}$, is defined as follows. The A-marginal of the contingency table is given by the marginal counts

$$n(\mathbf{x}_A) = \sum_{\mathbf{x}' \in \mathcal{X}:\, \mathbf{x}'_A = \mathbf{x}_A} n(\mathbf{x}') = \sum_{\mathbf{y}_{V \setminus A} \in \mathcal{X}_{V \setminus A}} n(\mathbf{x}_A, \mathbf{y}_{V \setminus A}), \quad \mathbf{x}_A \in \mathcal{X}_A = \times_{i \in A} \mathcal{X}_i,$$

i.e., the variables in $V \setminus A$ are 'summed out'. So if $|A| = k$, then these A-marginal counts form a k-dimensional contingency table of $\prod_{i \in A} r_i$ cells, and there are $\binom{d}{k}$ possible k-dimensional marginals ($k = 1, \ldots, d$). Likewise, the A-marginal distribution of the $\{p(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ distribution is defined by

$$p_A(\mathbf{x}_A) = \sum_{\mathbf{x}' \in \mathcal{X}:\, \mathbf{x}'_A = \mathbf{x}_A} p(\mathbf{x}') = \sum_{\mathbf{y}_{V \setminus A} \in \mathcal{X}_{V \setminus A}} p(\mathbf{x}_A, \mathbf{y}_{V \setminus A}), \quad \mathbf{x}_A \in \mathcal{X}_A.$$

Given the set $\Gamma = \{A : A \subset V\}$, called *generating class*, we define the following *log-linear model*:

$$\ln p(\mathbf{x}) = f_0 + \sum_{A \in \Gamma} f_A(\mathbf{x}_A), \tag{4}$$

where the individual terms represent interactions ($f_A : \mathcal{X}_A \to \mathbb{R}$ functions) corresponding to $A \in \Gamma$, for they depend on $\mathbf{x}$ only through $\mathbf{x}_A$, and the constant

term f_0 corresponds to $\emptyset \in \Gamma$ (it also fits into the forthcoming hierarchical structure of Γ). This is also in accord with the notation of the Gibbs field, see Section 1.2, where $|f_A|$'s are the energies of the configurations that correspond to the vertex-subsets in Γ of G. We will see that the log-linear model defines an MRF if and only if the generating class Γ consists of the cliques of G and of the subsets of them, see Section 2.2.

To meet some compatibility constraints, we consider *hierarchical* log-linear models: with any $A \in \Gamma$ and $A' \subset A$, the relation $A' \in \Gamma$ also holds, and some normalizing conditions are also needed (see [6]). If $\mathbb{P}$ obeys a hierarchical log-linear model, it means that it can be constructed as the product of functions defined on its lower dimensional margins up to a certain dimension. So it suffices to keep only the maximal interaction sets in Γ; such a Γ will be called *generating class* of the log-linear model.

In special hierarchical log-linear models (we will call them graphical), the generating class is specified with the set of maximal interactions

$$\mathcal{C} = \{C : C \text{ is a clique of the underlying graph}\},$$

and so, Γ consists of the complete subgraphs of the underlying graph. In this case, there is another equivalent form of Equation (4) that uses an exponential parametrization and shows that we are in *exponential family*:

$$p(\mathbf{x}) = \exp\left\{\sum_{C\in\mathcal{C}} \langle \theta_C, I_C(\mathbf{x}_C)\rangle - Z(\theta)\right\}.$$

Here $\theta = \{\theta_C : C \in \mathcal{C}\}$ is the canonical parameter, where

$$\theta_C = \{\theta_{C,\mathbf{y}_C}\,,\, \mathbf{y}_C \in \mathcal{X}_C\} \in \mathbb{R}^{|\mathcal{X}_C|}$$

is a vector, and so, θ is a $\sum_{C\in\mathcal{C}} |\mathcal{X}_C|$-dimensional vector, which dimension is usually less than $|\mathcal{X}| = \prod_{i=1}^n |\mathcal{X}_i|$. The canonical statistic I_C also takes on values in $\mathbb{R}^{|\mathcal{X}_C|}$ for every $C \in \mathcal{C}$. In fact, the I_C's are multiple indicator functions consisting of usual 0-1 indicator functions of all possible states in $\mathcal{X}_C$ (cells with coordinates in C). More exactly,

$$I_C = \{I_{C,\mathbf{y}_C}\,,\, \mathbf{y}_C \in \mathcal{X}_C\} \in \mathbb{R}^{|\mathcal{X}_C|},$$

where the usual indicator function $I_{C,\mathbf{y}_C}(\mathbf{x}_C)$ is 1 if $\mathbf{x}_C = \mathbf{y}_C$ and 0, otherwise. Further, $\langle .,. \rangle$ denotes the inner product in the above finite-dimensional spaces, and $Z(\theta)$ is the log-partition function (it does not depend on $\mathbf{x} \in \mathcal{X}$). In accord with Equation (4), $f_C = \theta_C I_C$.

In exponential family, the sums of the canonical statistics through an iid sample $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)} \in \mathcal{X}$, are the sufficient statistics entering into the parameter estimation. So based on this sample, the frequencies $n(\mathbf{x}_C)$'s of the cells within the cliques are the sufficient statistics. The mean value parameters (in other words, moment parameters) are their expectations: $m(\mathbf{x}_C) = np(\mathbf{x}_c)$. In regular exponential families, there is a one-to-one correspondence between the mean value and the canonical parameters, see [26]. Further, the ML-estimate of the mean value parameter comes from the moment-matching equations

$$m(\mathbf{x}_C) = n(\mathbf{x}_C), \quad C \in \mathcal{C}, \quad \mathbf{x} \in \mathcal{X}.$$

This system of equations is solved by the *Iterative Proportional Scaling* (IPS) algorithm of Section 2.5.

2.2 Graphical and decomposable models

In many applications we have a contingency table of large size: even in case of binary variables, there are 2^d cells the number of which grows exponentially with the number of variables d. Then the IPS algorithm of Section 2.5, going through the cells several times, is time-consuming. However, there are models, where the ML-estimate of the cell probabilities under the model's assumptions can be given by explicit formulas. These models are characterized by the special dependency structure of the variables when we build a graph or hypergraph on them. These are the decomposable models.

From now on, our log-lineal model is hierarchical, and therefore, we keep only the maximal interactions in Γ. Recall that we called this Γ the *generating class* of the model. Further, we assume that each variable is included in at least one interaction; in other words, all main effects are present. hhIn case of a special structure of the generating class, we can introduce an exact algorithm that goes through the $A \in \Gamma$ sets in a definite order, see the belief propagation algorithm of Section 2.5.

To discuss this, hypergraph notions may be used as follows. The generating class Γ uniquely defines the following hypergraph H: the vertices correspond to the variables and constitute the set $V = \{1, \ldots, d\}$, while the hyperedges are the elements of Γ (they are the maximal interaction sets). With our former assumption, each vertex is contained in at least one hyperedge. As the model is hierarchical, the subsets of the maximal interaction set are also interactions, but they are not hyperedges in H.

The *interaction graph* $G = G(H)$ corresponding to H, or equivalently, to the hierarchical log-linear model with generating class Γ, is defined in the following way. Its vertex set is again V, while the edges are as follows:

$$i \sim j \Leftrightarrow \{i, j\} \subseteq A \quad \text{for some} \quad A \in \Gamma,$$

i.e., two vertices are connected if and only if they are contained together in some interaction set.

The *clique hypergraph* H of a graph G (both are defined on the same vertex set) consists of hyperedges which are exactly the cliques of G. With another wording (see [23]), H is *conformal* (with G).

Observe that different connected components of the so-called interaction graph correspond to variables that are mutually (marginally) independent. Also note that different hierarchical models may have the same interaction graph, see the examples below. However, we introduce a class of models when there is a one-to-one correspondence between the model and its interaction graph. Therefore, the interaction graph is capable to describe such a model. To make it precise, we need some further definitions.

Definition 1 *The hierarchical log-linear model with generating class Γ is **graphical** if the hypergraph H defined above (with the hyperedges as the entries of the generating class Γ) is identical to the clique hypergraph of its interaction graph (see [17]), i.e., H is conformal (with G).*

Note that equivalently the definition means: the hierarchical log-linear model

with generating class Γ is **graphical** if the generating class Γ is identical to the cliques of its interaction graph.

For example, when the generating class is

$$\Gamma = \{\{1,2\},\{2,3\},\{1,3\}\}, \tag{5}$$

then the interaction graph has the clique $\{1,2,3\}$, which is not an interaction set. So our log-linear model is not a graphical interaction model. However, when the generating class is

$$\Gamma' = \{\{1,2,3\}\}, \tag{6}$$

then the interaction graph has the clique $\{1,2,3\}$, so our log-linear model is a graphical interaction model. Note that model (5) corresponds to the Ising model on 3 vertices. When there are more than 3 vertices, then, for example, a squared grid defines an Ising model, where the cliques are indeed the vertex-pairs, so those constitute the generating class at the same time.

Theorem 1 (see [17]) *The distribution $\mathbb{P}$ obeying the hierarchical log-linear model with generating class Γ defines an MRF (satisfies conditions (UF), (UG), (UL), and (UP)), if and only if the log-linear model is graphical (again, the cliques of the interaction graph G correspond to subsets of variables which are in interaction with each other).*

Now we investigate special graphical models, the decomposable ones.

Definition 2 *The hierarchical log-linear model with generating class Γ is **decomposable** if its interaction graph is decomposable.*

The definition of the (weak) decomposability of a graph is recursive.

Definition 3 *The graph G is decomposable if it is either a complete graph or its vertex-set V can be partitioned into disjoint vertex-subsets A, B, C such that*

- *C defines a complete subgraph;*
- *C separates A from B (in other words, C is a vertex cutset between A and B);*
- *the subgraphs generated by $A \cup C$ and $B \cup C$ are both decomposable.*

In this way, decomposability goes through to the logliner model as follows.

Proposition 2 *The log-linear model is decomposable if and only if the generating class Γ either consists of one set (G is the complete graph) or it is the disjoint union of the decomposable generating classes Γ_1 and Γ_2 (they contain no sets in common) such that there exist $A^* \in \Gamma_1$ and $B^* \in \Gamma_2$ with the following property:*

$$(\cup_{A\in\Gamma_1} A) \cap (\cup_{B\in\Gamma_2} B) = A^* \cap B^*.$$

It is important that decomposable models are subclasses of the graphical ones.

Proposition 3 (see [17], Corollary 7.5) *A log-linear model is graphical whenever it is decomposable.*

So, in case of contingency tables, the graphical interaction models (the cliques constitute the generating class) coincide with the MRF's. However, the decomposable models are proper subsets of them. In [6], the authors show examples of graphical interaction models that are not decomposable. We will cite some of these examples at the end of Section 2.3, after we have learned some equivalent notions of decomposability.

Note that some authors call the decomposable models Markov, as here the chain of the cliques behaves like a Markov chain, see Equation (8) in the subsequent Section 2.3. It is misleading, and again, the Markov chain property of the decomposable models is stronger than the condition for being an MRF.

2.3 Junction tree

If we have a graphical hierarchical log-linear model and the model is also decomposable, we can find a junction tree structure of the cliques by the following equivalences. Recall that under clique we understand a maximal complete subgraph (as in the statistics literature). Here we establish many equivalent properties of a decomposable graph, based on Proposition 2.5 of [14], Proposition 4 of [28], and the last section of [25]. For simplicity, the necessary notions are defined at the very place where they are introduced. Further explanation together with algorithmic aspects is to be found in Section 2.4.

Proposition 4 *The following properties are equivalent to the fact that G is decomposable:*

- *G is **triangulated** (with other words, **chordal**), i.e., every cycle in G of length at least four has a chord.*
- *G has a **perfect numbering** of its vertices such that in this labeling,*

$$\mathrm{bd}(i) \cap \{i+1, \ldots, d\} \tag{7}$$

  *is a complete subgraph, $i = 1, \ldots, d$. It is also called **single vertex elimination ordering** (see [25]), and obtainable with the Maximal Cardinality Search (MCS) algorithm of [23], see Section 2.4.*
- *G has the following **running intersection property (RIP)**: we can number the cliques of it to form a so-called **perfect sequence** $C_1, \ldots, C_k$ where each combination of the subgraphs induced by $H_{j-1} = C_1 \cup \cdots \cup C_{j-1}$ and C_j is a decomposition ($j = 2, \ldots, k$), i.e., the necessarily complete subgraph $S_j = H_{j-1} \cap C_j$ is a separator. More precisely, S_j is a vertex cutset between the disjoint vertex subsets $H_{j-1} \setminus S_j$ and $R_j = C_j \setminus S_j = H_j \setminus H_{j-1}$. This sequence of cliques is also called a **junction tree (JT)**. Here any clique C_j is the disjoint union of R_j (called **residual**), the vertices of which are not contained in any C_i, $i < j$ and of S_j (called **separator**) with the following property: there is an $i^* \in \{1, \ldots, j-1\}$ such that*

$$S_j = C_j \cap (\cup_{i=1}^{j-1} C_i) = C_j \cap C_{i^*}.$$

This (not necessarily unique) C_{i^*} is called parent clique of C_j. Here $S_1 = \emptyset$ and $R_1 = C_1$. Furthermore, if such an ordering is possible, a version may be found in which any prescribed set is the first one (see [21]).

Also equivalently, any path between C_i and C_j $(i \neq j)$ contains $C_i \cap C_j$.

Note that the junction tree is indeed a tree with vertices $C_1, \ldots, C_k$ and one less edges, that are the separators $S_2, \ldots, S_k$.

- ***Sundberg's criterion**: We can number the cliques of G as $C_1, \ldots, C_k$, where each combination of the subgraphs induced by $H_{j+1} = C_{j+1} \cup \cdots \cup C_k$ and C_j is a decomposition $(j = 1, \ldots, k-1)$, i.e., the necessarily complete subgraph $S_j = H_{j+1} \cap C_j$ is a separator. More precisely, S_j is a vertex cutset between the disjoint vertex subsets $H_{j+1} \setminus S_j$ and $R_j := C_j \setminus S_j$. Here R_j is called residual, and so $C_j = S_j \cup R_j$ is a disjoint union. This sequence of cliques forms the junction tree in the reversed RIP ordering.*

  So each C_j can be composed of one set of elements (R_j) which are missing in all C_i, for $i > j$ and one set $S_j = C_j \cap \cup_{i=j+1}^{k} C_i$ which is contained in some C_{i^*}, $i^* > j$. This (not necessarily unique) C_{i^*} is the former parent clique of C_j. Here $S_k = \emptyset$ and $R_k = C_k$.

  Furthermore, if such an ordering is possible, a version may be found in which any prescribed set is the last one (see [21]). As the Sundberg's ordering of the cliques is opposite to the RIP ordering, in the RIP ordering any prescribed clique can be the first one.

- *G is **recursively simplicial**, see [25]. A non-empty graph G is recursively simplicial if it contains a simplicial vertex, and when that is removed, any graph that remains is recursively simplicial.*

  *A vertex is called **simplicial** in a graph if its neighbors form a complete subgraph. Every decomposable graph with at least two vertices has at least two simplicial vertices; if the graph is not complete, these vertices can be chosen to be non-adjacent (see [25]). We can arrange that C_1 and C_k contain the simplicial vertices.*

- *There is a labeling of the vertices such that the adjacency matrix contains a **reducible zero pattern (RZP)**. It means that the zero entries in the upper-diagonal part of the adjacency matrix form an index set that is reducible in the following sense. The index set I, which is the subset of the set of edges $\{(i, j) : 1 \leq i < j \leq d\}$, is called reducible if for each $(i, j) \in I$ and $h = 1, \ldots, i-1$, we have $(h, i) \in I$ or $(h, j) \in I$ or both.*

  Indeed, this convenient labeling is a perfect numbering (7) of the vertices.

Note that decomposable graphs are special perfect graphs. A perfect graph is defined as follows: it and any spanning subgraph of it has the same chromatic number as the size of the maximum clique (maximum size maximal clique in the graph or in the spanning subgraph in question). L. Lovász (1972) proved that the complement of a perfect graph is also perfect. It can also be proven that a graph is perfect if and only if any odd cycle (of length greater than three) of it and (in view of the above) of its complement has a chord, see the related exercises of [15].

Also observe that in the RIP ordering, the cliques also form a *Markov chain*: the conditional distribution of $\mathbf{X}_{R_j}$ conditioned on $\mathbf{X}_{C_1\cup\cdots\cup C_{j-1}}$ is the conditional distribution of $\mathbf{X}_{R_j}$ conditioned on $\mathbf{X}_{S_j}$, i.e.,

$$p(\mathbf{x}_{R_j} \,|\, \mathbf{x}_{C_1\cup\cdots\cup C_{j-1}}) = p(\mathbf{x}_{R_j} \,|\, \mathbf{x}_{S_j}), \tag{8}$$

where p is the pmf in discrete, and pdf in continuous cases. Sometimes this is called Markovity, but again, it is stronger than being an MRF.

In decomposable log-linear models, there are also exact ML-estimates for the cell probabilities, and so, for the moment parameters. Here we cite some results of [14]. If we have the RIP ordering $C_1, \ldots, C_k$ of the cliques with separators $S_2, \ldots, S_k$, then for the cell probabilities we have

$$p(\mathbf{x}) = \frac{\prod_{j=1}^{k} p(\mathbf{x}_{C_j})}{\prod_{j=2}^{k} p(\mathbf{x}_{S_j})} = \frac{\prod_{C\in\mathcal{C}} p(\mathbf{x}_C)}{\prod_{S\in\mathcal{S}} p(\mathbf{x}_S)^{\nu(S)}}, \quad \mathbf{x}\in\mathcal{X} \tag{9}$$

where $\mathcal{C}$ is the set of the cliques, $\mathcal{S}$ is the set of the separators along the JT, and $\nu(S)$ is the multiplicity of the occurrence of the separator S in the above JT of G. Hence, the ML-estimate of the mean vector is given by the explicit formula

$$\hat{m}(\mathbf{x}) = \frac{\prod_{j=1}^{k} n(\mathbf{x}_{C_j})}{\prod_{j=2}^{k} n(\mathbf{x}_{S_j})} = \frac{\prod_{C\in\boldsymbol{C}} n(\mathbf{x}_C)}{\prod_{S\in\boldsymbol{S}} n(\mathbf{x}_S)^{\nu(S)}}, \quad \mathbf{x}\in\mathcal{X} \tag{10}$$

and that of the cell probabilities is $\hat{p}(\mathbf{x}) = \frac{\hat{m}(\mathbf{x})}{n}$, $\mathbf{x}\in\mathcal{X}$.

Equation (9) also induces the following factorization:

$$p(\mathbf{x}) = \prod_{i=1}^{k} p(\mathbf{x}_{R_j}|\mathbf{x}_{S_j}) \tag{11}$$

in the RIP ordering, where R_j's are the residuals, S_j's are the separators of the cliques $C_1, \ldots, C_k$; further, $R_1 = C_1$, $S_1 = \emptyset$, so $p(\mathbf{x}_{R_1}|\mathbf{x}_{S_1}) = p(\mathbf{x}_{C_1})$. It is in accord with Equation (8) and also resembles the factorization (DF) of a directed graph. It also gives a possible factorization in (UF), where $Z = 1$ and the compatibility function is, in fact, a conditional probability of the clique's residual on the clique's separator: $\Psi_{C_j}(\mathbf{x}_{C_j}) = p(\mathbf{x}_{R_j}|\mathbf{x}_{S_j})$ for any clique $C_j = R_j \cup S_j$ and any state configuration $\mathbf{x}_{C_j}$ within it.

Figure 1 illustrates that decomposable models are subclasses of the graphical ones. Figure 1(a) is the complete graph corresponding to the unrestricted model, while 1(b) corresponds to the model $X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_5 \,|\, \{X_3, X_4\}$, in terms of conditional independences; (c) and (d) are not decomposable: (c) corresponds to the model $X_1 \perp\!\!\!\perp X_4 \perp\!\!\!\perp X_5 \,|\, \{X_2, X_3\}$ and vice versa, $X_2 \perp\!\!\!\perp X_3 \,|\, \{X_1, X_4, X_5\}$, while (d) corresponds to the model $X_1 \perp\!\!\!\perp X_3 \,|\, \{X_2, X_4, X_5\}$ and $X_2 \perp\!\!\!\perp X_4 \,|\, \{X_1, X_3, X_5\}$. In (a) and (b), the cliques constitute a junction tree in the Sundberg's ordering (so they both are decomposable), and Equation (9) is applicable for their factorization. However, in (c) we can use the factorization according to the log-linear model. Actually, its equivalent form for the log-probabilities is as follows:

$$\ln p(x_1, x_2, x_3, x_4, x_5) = f_{1,2}(x_1, x_2) + f_{1,3}(x_1, x_3) + f_{2,4}(x_2, x_4) + f_{2,5}(x_2, x_5) + f_{3,4}(x_3, x_4) + f_{3,5}(x_3, x_5) + f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4) + f_5(x_5) + f_0.$$

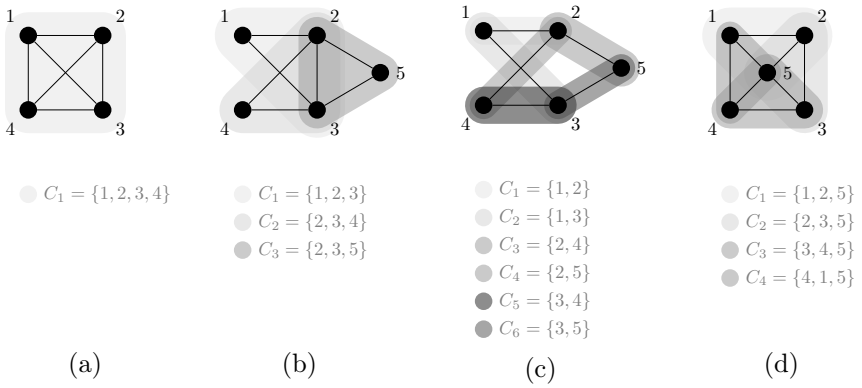

Figure 1: Examples of graphical models: (a) and (b) are also decomposable; (c) and (d) are not. One may think that (d) is triangulated, but it is not: 1-2-3-4-1 is a chordless 4-cycle in it.

In the following example, let us consider the rv's X_1, X_2, X_3, and assume that X_2 and X_3 are independent conditioned on X_1. It means that the generating class of the log-linear model is

$$\Gamma = \{\{1, 2\}, \{1, 3\}\}, \tag{12}$$

and the interaction graph has the cliques $\{1, 2\}$ and $\{1, 3\}$. This log-linear model is decomposable with the only separator $S = \{1\}$ between the cliques. From Equation (9), we get the formula

$$p(x_1, x_2, x_3) = \frac{p(x_1, x_2)p(x_1, x_3)}{p(x_1)} = p(x_1, x_2)p(x_3|x_1) \tag{13}$$

which gives possible factorizations. The graph here was $2 - 1 - 3$.

As the last example, consider the graph of Figure 2, where the generating class of the log-linear model is

$$\Gamma = \{\{1, 3\}, \{2, 3\}, \{3, 4\}, \{4, 5, 6\}\}. \tag{14}$$

As the entries of Γ are the cliques of the interaction graph, this log-linear model is a graphical interaction model, and it is also decomposable with the cliques $\{1, 3\}, \{2, 3\}, \{3, 4\}, \{4, 5, 6\}$, which form a junction tree in Sundberg's ordering with the separators $\{3\}, \{3\}, \{4\}$. Therefore, the probabilities in this model can be decomposed as

$$p(x_1, x_2, x_3, x_4, x_5, x_6) = \frac{p(x_1, x_3) \cdot p(x_2, x_3) \cdot p(x_3, x_4) \cdot p(x_4, x_5, x_6)}{p^2(x_3) \cdot p(x_4)}$$

for all $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6) \in \mathcal{X}$.

2.4 Numerical algorithms to find a junction tree

To find the structure, where one of the equivalent criteria (e.g., triangulatedness) of Proposition 4 holds, we can use the MCS (Maximal Cardinality Search)

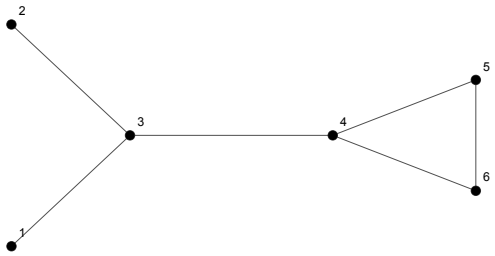

Figure 2: Interaction graph with cliques in (14).

method of [23]. For a simple version of MCS see the pseudocode of [11], page 312.

The simple MCS gives label d to an arbitrary vertex. Then labels the vertices consecutively, from d down to 1, choosing as the next to label a vertex with a maximum number of previously labeled neighbors and breaks ties arbitrarily. Note that [14] labels the vertices conversely, and so, there our perfect labeling (7) happens in the opposite direction. The MCS ordering is far not unique, and this simple version is not always capable to find the JT structure behind a triangulated graph in one run, but another run is needed as follows.

To get the cliques of the triangulated graph and construct a JT structure based on the ordering $d, d-1, \ldots, 2, 1$, provided by the MCS, we proceed as follows. Take the vertex labeled i and its higher labeled neighbors:

$$M_i := \{i\} \cup \{j : j > i \text{ and } j \sim i\}, \qquad i = 1, \ldots, d.$$

Since the input was a perfect ordering, every M_i will be complete subgraph. Then $C_1, \ldots, C_k$ is obtained by deleting all M_i's that are subsets of another one, that is by keeping the maximal complete subgraphs (the cliques) of them. If we label the cliques consecutively, we obtain the Sundberg's ordering; whereas, the reversed ordering gives the RIP ordering of them.

Note that, in this way, we are able to get a new labeling of the vertices by partitioning them according to the JT structure. Let us form the separators and residuals of the cliques in the Sundberg's ordering. Then vertices can be relabeled by permuting them between the separators and the residuals, and within the residuals, see Figure 3. This relabeling will not hurt the JT structure and it is another perfect ordering of the vertices.

It gives us a more causal way to look at the vertices. For example, we can form a DAG in this modified perfect ordering on the skeleton of G such that $j \to i$ if $i < j$ and $i \sim j$. Note that, in the above relabeling, a strict perfect numbering of the vertices is obtained, in which

$$\mathrm{bd}(i) \cap \{1, \ldots, i-1\}$$

is a complete subgraph, for $i = 1, \ldots, d$. This indicates that the reversed labeling of the vertices is perfect too.

We remark that for a moderate number of vertices, we can as well proceed as follows. We start with a simplicial vertex and run the MCS with the restriction that first we exhaust the cliques. We can do so by counting the degrees of

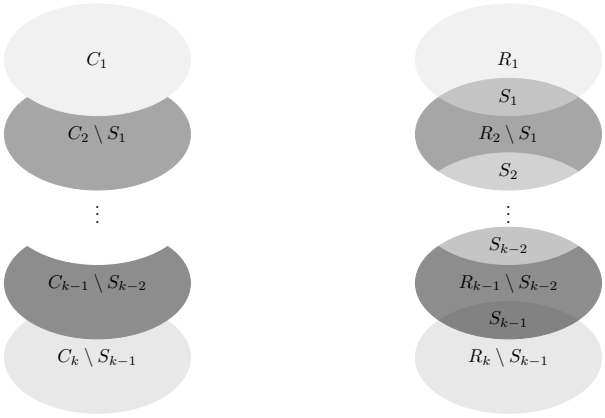

Figure 3: Relabeling the variables within the junction tree in the Sundberg's ordering

the vertices which have the same number of formerly labeled neighbors. Starting from vertex labeled d, first we select vertices with the same degree. At the moment, when there are no more such vertices, the first residual (R_1 in the RIP ordering) is exhausted. Next come the vertices with higher degree, which belong to the first separator (S_2) and have neighbors in other cliques too. At the moment, when the degrees start again decrease, the first clique (C_1) is exhausted, etc. This method gives the cliques and the separators in the RIP ordering that will be used by the so-called belief propagation algorithm of Section 2.5. However, this method is not the best as for the computational complexity. Numerical issues are discussed in Tarjan, Yannakakis [23] together with other issues, for example the lexicographical labeling and the fill-in procedure for undirected graphs which are not triangulated.

Note that fill-in can be also defined for directed graphs, see e.g., [20]. Recall that to have Markov equivalences, the graph has to be first moralized, and this undirected moral graph is then triangulated, see also Section 4.1. The process of triangulation is called fill-in in [13] and [25] too.

It can easily be seen that in a chordal graph, the perfect ordering is also a suitable ordering of the vertices in which the adjacency matrix has the reducible zero pattern I. Note that by Proposition 6 of [28], the cliques (in the reversed RIP ordering, that is in the Sundberg's ordering) can be obtained as discussed before:

$$M_i := \{i\} \cup \{j : j > i \text{ and } (i,j) \notin I\}, \qquad i = 1, \ldots, k,$$

then $C_1, \ldots, C_k$ is again obtained by deleting all M_i's that are subsets of another one. A possible MCS-ordering and JT structure is shown in Figure 4.

Another construction of a JT from a so-called cluster tree is as follows (see [25]). Call the cliques clusters, and first have all separators (intersections) between the cluster pairs.

The so obtained cluster graph, with vertices as the clusters and edges as the separators, usually contains cycles. Then find the *maximal weight spanning tree* of this cluster graph with usual algorithms of Kruskal, Prim, see, e.g., on

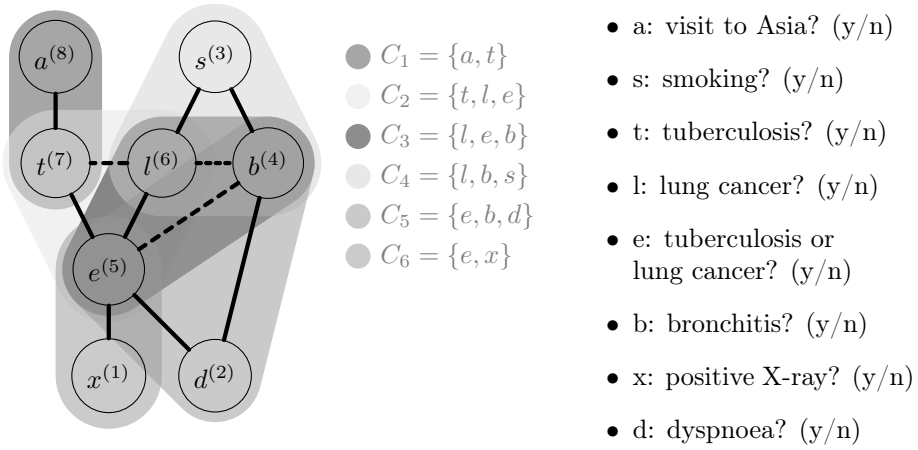

Figure 4: The stylized example of [13], where the set of variables $V = \{a, s, t, l, b, e, x, d\}$ is labeled (see the superscript) based on MCS, starting from vertex a of label 8. (It is a backward numbering, so the perfect elimination ordering is x, d, s, b, e, l, t, a.) On the left panel see the clique structure of the triangulated graph in the RIP ordering.

page 1147 of [11]. Again, the vertices are the cliques, while the edges are the separators with weights that are equal to their size (cardinality). Any maximal weight spanning tree (there can be more than one) will be a computationally economic JT. More exactly, Proposition 4 of [25] states that any maximal weight spanning tree of the cardinality weighted clique graph is a junction tree for the original graph.

If we have the joint distribution, and we want to find a tree-structured graph that defines an MRF over it, then we can use the Chow–Liu algorithm of [3]. Based on the empirical probabilities (estimated from the sample) of the vertices and vertex-pairs (edges), the likelihood of the spanning tree over the vertices is maximized with information theoretical tools, see [25]. More general structures are investigated in [22].

If the variables are binary, and the interactions are singletons or pairwise, the min-cut algorithm can also be applied, which runs in polynomial time, see [11]. For this purpose we construct an edge-weighted graph, the cuts of which to be minimized are just the energies in the Gibbs model. The vertices are either in spin state 0 or 1, and the weights are obtained from the exponents of the potential functions, see [1] for details.

2.5 Iterative scaling, belief propagation, and mode prediction

In hierarchical log-linear models, the mean value parameters, and so, the cell probabilities are estimated based on the clique frequencies, and are obtainable by the IPS (Iterative Proportional Scaling) algorithm, see [25] (page 97) and [14] (page 82). In the heart of this algorithm lies the following: we want to make the clique probabilities equal to the corresponding relative frequencies, for all

cliques. Recall that

$$\{n(\mathbf{x}_C), \quad \mathbf{x}_C \in \mathcal{X}_C, \quad C \in \mathcal{C}\}$$

is a sufficient statistic for the canonical parameters of the log-linear model. Moreover, as we are in exponential family, the C-marginals of the ML-estimate $\hat{m}$ of the mean value parameter m satisfy the system of equations

$$m(\mathbf{x}_C) = n(\mathbf{x}_C), \quad C \in \mathcal{C}, \quad \mathbf{x} \in \mathcal{X}$$

for all the cliques (and consequently, for their subsets), but not for larger subsets of the vertices. To solve the above system, we will recursively adjust the above marginal counts going through each clique in a cyclic iteration that finds the fixed point of the mapping $T = T_{C_1} \dots T_{C_k}$ (if there are k cliques in $\mathcal{C}$), where

$$T_{C_i} m(\mathbf{x}) = m(\mathbf{x}) \frac{n(\mathbf{x}_{C_i})}{m(\mathbf{x}_{C_i})}, \quad i = 1, \dots, k.$$

Note that $T_{C_i} m(\mathbf{x}_{C_i}) = n(\mathbf{x}_{C_i})$ and $\sum_{\mathbf{x} \in \mathcal{X}} T_{C_i} m(\mathbf{x}) = n$.

So starting from some $m^{(0)}$, the iteration is

$$m^{(t)}(\mathbf{x}) = T m^{(t-1)}(\mathbf{x}), \quad \mathbf{x} \in \mathcal{X}.$$

In [17] it is proved that if $n(\mathbf{x}_C) > 0$ and $m^{(0)}(\mathbf{x}_C) > 0$, $\forall \mathbf{x} \in \mathcal{X}$ and $\forall C \in \mathcal{C}$, then the sequence $m^{(t)}(\mathbf{x})$ converges as $t \to \infty$, for all $\mathbf{x} \in \mathcal{X}$. With some additional condition, namely, that $m^{(0)}(\mathbf{x}_A) = n(\mathbf{x}_A)$ cannot hold for $A \notin \mathcal{C}$, the sequence $m^{(t)}(\mathbf{x})$ converges to the theoretically guaranteed unique ML estimate of m:

$$m^{(t)}(\mathbf{x}) \to \hat{m}(\mathbf{x}) \quad \text{as} \quad t \to \infty, \quad \forall \mathbf{x} \in \mathcal{X}$$

or equivalently, $\frac{m^{(t)}(\mathbf{x})}{n} \to \hat{p}(\mathbf{x})$. The proof is based on information divergence minimization, see [17, 25]. The additional condition excludes the possibility that some extra subset of variables is added to the prescribed set of interactions (which are the cliques). In particular, the cell frequencies do not provide a good starting, as they belong to the saturated model. The suggested starting is the uniform distribution over the cells, i.e., $m^{(0)}(\mathbf{x}) = \frac{n}{c}$, where $c = |\mathcal{X}|$ is the total number of the cells.

Note that the same idea is hidden behind the so-called covariance selection method in the Gaussian case, see Section 3.3.

In general, in hierarchical log-linear models, we cannot give the ML estimate of the mean value parameter in explicit form, this is why the infinite iteration of IPS is needed that converges to this estimate. However, when the log-linear model is decomposable, we have the ML estimate in explicit form, see (10), and in accord with this, we can construct an iteration that converges in two runs. The iteration facilitates the quick computation of the clique marginals. Here the special structure of the cliques and separators is exploited.

For the cliques of the junction tree (in the reversed RIP ordering), together with separators, we apply the so-called *message-passing*, in other words, *belief propagation* algorithm so that we update their potentials in such a way, that at the end, they become the clique marginals. Let A and B be two consecutive

cliques, and S be the separator between them. Starting with some potentials (see later), and denoting by * the newly updated potential, the algorithm is:

$$\begin{aligned}
\psi^*_S(\mathbf{x}_S) &= \sum_{\mathbf{x}_{A\setminus S}\in\mathcal{X}_{A\setminus S}} \psi_A(\mathbf{x}_S,\mathbf{x}_{A\setminus S}), && \forall \mathbf{x}_S\in\mathcal{X}_S \\
\psi^*_B(\mathbf{x}_B) &= \psi_B(\mathbf{x}_B)\cdot\frac{\psi^*_S(\mathbf{x}_S)}{\psi_S(\mathbf{x}_S)}, && \forall \mathbf{x}_B\in\mathcal{X}_B \\
\psi^{**}_S(\mathbf{x}_S) &= \sum_{\mathbf{x}_{B\setminus S}\in\mathcal{X}_{B\setminus S}} \psi^*_B(\mathbf{x}_S,\mathbf{x}_{B\setminus S}), && \forall \mathbf{x}_S\in\mathcal{X}_S \\
\psi^*_A(\mathbf{x}_A) &= \psi_A(\mathbf{x}_A)\cdot\frac{\psi^{**}_S(\mathbf{x}_S)}{\psi^*_S(\mathbf{x}_S)}, && \forall \mathbf{x}_A\in\mathcal{X}_A.
\end{aligned} \tag{15}$$

These equations hold for any state-configurations $\mathbf{x}_A, \mathbf{x}_S, \mathbf{x}_B$ within the cliques.

Algorithm (15) can be thought of as the so-called *sum-product algorithm* that gives the following:

$$\begin{aligned}
\sum_{\mathbf{y}\in\mathcal{X}_{A\setminus S}} \psi^*_A(\mathbf{x}_S,\mathbf{y}) &= \sum_{\mathbf{y}\in\mathcal{X}_{A\setminus S}} \psi_A(\mathbf{x}_S,\mathbf{y})\frac{\psi^{**}_S(\mathbf{x}_S)}{\psi^*_S(\mathbf{x}_S)} = \frac{\psi^{**}_S(\mathbf{x}_S)}{\psi^*_S(\mathbf{x}_S)}\sum_{\mathbf{y}\in\mathcal{X}_{A\setminus S}}\psi_A(\mathbf{x}_S,\mathbf{y}) \\
&= \frac{\psi^{**}_S(\mathbf{x}_S)}{\psi^*_S(\mathbf{x}_S)}\psi^*_S(\mathbf{x}_S) = \psi^{**}_S(\mathbf{x}_S) = \sum_{\mathbf{y}\in\mathcal{X}_{B\setminus S}}\psi^*_B(\mathbf{x}_S,\mathbf{y}).
\end{aligned}$$

So $\sum_{\mathbf{y}\in\mathcal{X}_{B\setminus S}}\psi^*_B(\mathbf{x}_S,\mathbf{y}) = \sum_{\mathbf{y}\in\mathcal{X}_{A\setminus S}}\psi^*_A(\mathbf{x}_S,\mathbf{y})$ after one back and forth step, which means local consistency, see [13] (page 181).

Start with clique potentials obtained from conditional probability tables, whereas the separator potentials can be constantly 1's. So ψ_C contains the product of marginal or conditional probabilities, affecting variables included in C, and such that the product of ψ_C's for $C\in\mathcal{C}$, with normalizing constant Z, gives the formula (UF). It means that the joint distribution is already factorized in some form with respect to the cliques. To find all clique and separator marginals, we first run the algorithm in the reversed RIP, that is, in the Sundberg's ordering $C_1,\ldots,C_k$ of the cliques. In this forward step we start at C_1 (called root), and via the separators, end at C_k. The so obtained potential of C_k is already the clique potential. To obtain all the clique potentials, we have to run the algorithm again, that is to make a backward step (in the RIP ordering). In this way, each separator appears with multiplicity in the calculations.

It is proven (see [12, 25]) that at the end, $\psi^{**}_{C_i}(\mathbf{x}_{C_i}) = p(\mathbf{x}_{C_i})$ and $\psi^{**}_{S_i}(\mathbf{x}_{S_i}) = p(\mathbf{x}_{S_i})$, $i=1,\ldots,k$; so the iteration leads to the clique marginals. In other wording, in the forward steps, the cliques collect the information from all of its neighbors (parent cliques on the JT) recursively; whereas, in the backward steps, they distribute the information to them. This is the so-called HUGIN version of the belief propagation algorithm, whereas the original version of [12] does not store the separator potentials.

Basically, we have historical data or empirical observation a priori. These provide us with so-called probability tables of conditional probabilities and also suggest the causal links. After, if a new observation comes in with some evidences (observed values of some of its variables), we can substitute those (or

make them absorbed by other cliques), and get probabilities of the other variables, see the expert system description of [13].

Sometimes, we want to take the mode of the log-linear distribution, i.e., to find the most probable state. In the junction tree framework, it is the *max-product algorithm* that does this. The max-product algorithm is practically the same as the sum-product algorithm (15) with the difference, that instead of summation we take maxima over the same sets.

For example, if we want to predict the mode (the most probable value of a variable) based on the observed values of the others, it suffices to consider the observed values of those variables which share cliques and/or separators with the target variable. For example, let $X_1, \ldots, X_d$ be categorical variables, where X_i takes on r_i distinct values. We want to predict the value of the target variable (say, X_1) based on the given values $x_2, \ldots, x_d$ of the others. If x_{1i} denotes the ith possible value of X_1, we are looking for the conditional probabilities

$$p(x_{1i}|x_2, \ldots, x_d) = \frac{p(x_{1i}, x_2, \ldots, x_d)}{p(x_2, \ldots, x_d)}, \quad i = 1, \ldots, r_1 \tag{16}$$

and find the i^* for which it is maximal. This is a discrete maximization (integer programming) task. Then x_{1i^*} is the mode of X_1 conditioned on the given values of the other variables, and this is our prediction for X_1. For example, if $X_2, \ldots, X_d$ are possible symptoms, and X_1 is the diagnosis, then x_{1i^*} is the most likely diagnosis under the given symptoms.

If $X_1, \ldots, X_d$ is a perfect numbering, and the variables are organized into a *junction tree* structure, in the possession of the clique (and separator) potentials (obtained by the belief propagation algorithm), we proceed as follows. We find

$$p(\mathbf{x}^i) := p(x_{1i}, x_2, \ldots, x_d) = \frac{\prod_C p(\mathbf{x}^i_C)}{\prod_S p(\mathbf{x}^i_S)}, \quad i = 1, \ldots, r_1.$$

However, the C's and S's that do not contain X_1 can be disregarded, as those marginal counts do not depend on i at all. Therefore,

$$p(\mathbf{x}^i) \propto q_i := \frac{\prod_{C:\, X_1 \in C} p(\mathbf{x}^i_C)}{\prod_{S:\, X_1 \in S} p(\mathbf{x}^i_S)}.$$

Eventually,

$$p(x_{1i}|x_2, \ldots, x_d) = \frac{q_i}{\sum_{j=1}^{r_1} q_j}, \quad i = 1, \ldots, r_1$$

and a discrete maximization in i closes the mode finding procedure for X_1.

Again, the important clique and separator marginals (where X_1 is included) are obtained through the junction tree iteration, which can be stopped at the desired place.

Note that the estimation process can be extended to directed graphs or to CG models, where some of the variables can be continuous (scaled). We can either categorize them or assuming, that they are Gaussian (conditioned on the discrete ones), similar procedures are available via covariance estimates, see Section 3.

3 Continuous MRF's and Gaussian graphical models

3.1 Partitioned Covariance Matrices and Partial Correlations

Here we consider the multivariate Gaussian distribution which is able to define so-called compositional graphoids (see [16]), and thus, embody the prototype of continuous multivariate distributions with existing second moments, where pairwise relations rule the joint distribution of the components.

Let $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a d-variate Gaussian random vector with expectation (vector) $\boldsymbol{\mu}$ and positive definite, symmetric $d \times d$ covariance matrix $\boldsymbol{\Sigma}$. Note that this distribution belongs to the exponential family with canonical parameter $(\boldsymbol{\Sigma}^{-1}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$. The also positive definite, symmetric matrix $\boldsymbol{\Sigma}^{-1}$ of entries σ^{ij} is called *concentration matrix*, and its zero entries indicate conditional independences between two components of $\mathbf{X}$, conditioned on the remaining components. This is supported by the following facts.

Proposition 5 *Let the $(p+q)\times(p+q)$ covariance matrix $\boldsymbol{\Sigma} > 0$ be partitioned as*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

where $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{22}$ are covariance matrices of $\mathbf{X}_1$ and $\mathbf{X}_2$, whereas $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^T$ is their cross-covariance matrix. Then the symmetric matrix $\boldsymbol{\Sigma}^{-1} > 0$ has the following partitioned form:

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_{1|2}^{-1} & -\boldsymbol{\Sigma}_{1|2}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{1|2}^{-1} & \boldsymbol{\Sigma}_{22}^{-1} + \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{1|2}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \end{pmatrix}, \tag{17}$$

where

$$\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

Further, $\boldsymbol{\Sigma} > 0$ is equivalent to the fact that both $\boldsymbol{\Sigma}_{22}$ and $\boldsymbol{\Sigma}_{1|2}$ are regular (invertible) matrices (actually, they are positive definite).

Theorem 2 *Let $(\mathbf{X}_1^T, \mathbf{X}_2^T)^T \sim \mathcal{N}_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a random vector, where the expectation $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ are partitioned (with block sizes p and q) in the following way:*

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Then the conditional distribution of the random vector $\mathbf{X}_1$ conditioned on $\mathbf{X}_2 = \mathbf{x}_2$ is $\mathcal{N}_p(\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) + \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{1|2})$ distribution.

Note that the conditional covariance matrix $\boldsymbol{\Sigma}_{1|2}$ does not depend on $\mathbf{x}_2$ of the condition. Further, for the conditional expectation, which is the expectation of the conditional distribution, we get that

$$\mathbb{E}(\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2) = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) + \boldsymbol{\mu}_1.$$

Therefore,

$$\mathbb{E}(\mathbf{X}_1|\mathbf{X}_2) = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2) + \boldsymbol{\mu}_1$$

which is a linear function of the coordinates of $\mathbf{X}_2$. In the $p = q = 1$ case, it is called *regression line*, while in the $p = 1$, $q > 1$ case, *regression plane*. Summarizing, in case of the multidimensional Gaussian distribution, the regression functions are linear functions of the variables in the condition, which fact has important consequences in the multivariate statistical analysis. Since $\boldsymbol{\mu}$ is just a shift, in the sequel, we will assume $\boldsymbol{\mu} = \mathbf{0}$, i.e., the variables are mean centered.

Theorem 3 *Let $\mathbf{X} = (X_1, \ldots, X_d)^T \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Sigma})$ be a random vector, and let $V := \{1, \ldots, d\}$ denote the index set of the variables, $d \geq 3$. Assume that $\boldsymbol{\Sigma}$ is positive definite. Then*

$$r_{X_iX_j|\mathbf{X}_{V\setminus\{i,j\}}} = \frac{-\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}} \qquad i \neq j,$$

where $r_{X_iX_j|\mathbf{X}_{V\setminus\{i,j\}}}$ denotes the partial correlation coefficient between X_i and X_j after eliminating the effect of the remaining variables $\mathbf{X}_{V\setminus\{i,j\}}$. Further,

$$\sigma^{ii} = 1/(\mathrm{Var}(X_i|\mathbf{X}_{V\setminus\{i\}}), \qquad i = 1, \ldots, d$$

is the reciprocal of the conditional (residual) variance of X_i conditioned on the other variables $\mathbf{X}_{V\setminus\{i\}}$.

Note that the $r_{X_iX_j|\mathbf{X}_{V\setminus\{i,j\}}} = 0 \iff \sigma^{ij} = 0$ equivalence can heuristically be explained as follows. It suffices to prove for the $i = 1$, $j = 2$ case. $r_{X_1X_2|\mathbf{X}_{V\setminus\{1,2\}}} = 0$ means that when regressing X_1 and X_2 with $\mathbf{X}_{V\setminus\{1,2\}}$, the residuals have 0 covariance. This is equivalent to that the residual covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ is diagonal, where $\boldsymbol{\Sigma}_{11}$ is the upper left 2×2 block of $\boldsymbol{\Sigma}$, and the other blocks are constructed accordingly. This in turn is equivalent to that the inverse of the residual covariance matrix is also diagonal, and this is just the upper left 2×2 block of $\boldsymbol{\Sigma}^{-1}$, see (17).

Definition 4 *Let $\mathbf{X} \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Sigma})$ be random vector with $\boldsymbol{\Sigma}$ positive definite. Consider the regression plane*

$$\mathbb{E}(X_i|\mathbf{X}_{V\setminus\{i\}} = \mathbf{x}_{V\setminus\{i\}}) = \sum_{j\in V\setminus\{i\}} \beta_{ji\cdot V\setminus\{i\}} x_j, \quad j \in V \setminus \{i\},$$

*where x_j's are the coordinates of $\mathbf{x}_{V\setminus\{i\}}$. Then we call the coefficient $\beta_{ji\cdot V\setminus\{i\}}$ the **partial regression coefficient** of X_j when regressing X_i with $\mathbf{X}_{V\setminus\{i\}}$, $j \in V \setminus \{i\}$.*

Theorem 4

$$\beta_{ji\cdot V\setminus\{i\}} = -\frac{\sigma^{ij}}{\sigma^{ii}}, \quad j \in V \setminus \{i\}.$$

Corollary 1 *An important consequence of Theorems 3 and 4 is that*

$$\beta_{ji\cdot V\setminus\{i\}} = r_{X_iX_j|\mathbf{X}_{V\setminus\{i,j\}}}\sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}} = r_{X_iX_j|\mathbf{X}_{V\setminus\{i,j\}}}\sqrt{\frac{\mathrm{Var}(X_i|\mathbf{X}_{V\setminus\{i\}})}{\mathrm{Var}(X_j|\mathbf{X}_{V\setminus\{j\}})}}, \quad j \in V\setminus\{i\}.$$

(The formula is analogous to the one of unconditioned regression.) So only the variables X_j's whose partial correlation with X_i (after eliminating the effect of the remaining variables) is not 0, enter into the regression of X_i with the other variables.

3.2 Testing hypotheses about partial correlations

For $i \neq j$ we want to test

$$H_0 \,:\, r_{X_i X_j | \mathbf{X}_{V\setminus\{i,j\}}} = 0,$$

i.e., that X_i and X_j are conditionally independent conditioned on the remaining variables. Equivalently, H_0 means that $\beta_{ij|V\setminus\{i\}} = 0$, $\beta_{ji|V\setminus\{j\}} = 0$, or simply, $\sigma^{ij} = \sigma^{ji} = 0$ ($\mathbf{\Sigma} > 0$ is assumed).

To test H_0 in some form, several exact tests are known that are usually based on likelihood ratio tests. The following test uses the empirical partial correlation coefficient, denoted by $\hat{r}_{X_i X_j | \mathbf{X}_{V\setminus\{i,j\}}}$, and the following statistic is based on it:

$$B = 1 - (\hat{r}_{X_i X_j | \mathbf{X}_{V\setminus\{i,j\}}})^2 = \frac{|\boldsymbol{S}_{V\setminus\{i,j\}}| \cdot |\boldsymbol{S}_V|}{|\boldsymbol{S}_{V\setminus\{i\}}| \cdot |\boldsymbol{S}_{V\setminus\{j\}}|},$$

where $\boldsymbol{S}$ is the sample size times the empirical covariance matrix of the variables in the subscript (its entries are the product-moments).

It can be proven that under H_0, the test statistic

$$t = \sqrt{n-d} \cdot \sqrt{\frac{1}{B} - 1} = \sqrt{n-d} \cdot \frac{\hat{r}_{X_i X_j | \mathbf{X}_{V\setminus\{i,j\}}}}{\sqrt{1 - (\hat{r}_{X_i X_j | \mathbf{X}_{V\setminus\{i,j\}}})^2}}$$

is distributed as Student's t with $n-d$ degrees of freedom. Therefore, we reject H_0 for large values of $|t|$.

3.3 The undirected model and the covariance selection

Let $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \mathbf{\Sigma})$ be a d-dimensional Gaussian random vector, and form a graph $\mathcal{G}$ on the vertex-set V, where V corresponds to the components of $\mathbf{X}$ and the edges are drawn according to the rule

$$i \sim j \Leftrightarrow \sigma^{ij} \neq 0, \quad i \neq j.$$

This is called *Gaussian graphical model*. For practical purposes we use the empirical partial correlation coefficients, and based on them, the above exact test to check whether they significantly differ from 0 or not. If we put zeros into the no-edge positions ij's of the inverse covariance matrix, we can fit a so-called *covariance selection model*. The restricted covariance matrix is denoted by $\mathbf{\Sigma}^*$.

With the help of the concentration matrix $\boldsymbol{K} = \mathbf{\Sigma}^{-1}$ and the vector $\mathbf{h} = \boldsymbol{K}\boldsymbol{\mu}$, the log-density of $\mathbf{X}$ has the following form:

$$\ln f(\mathbf{x}) = c - \frac{1}{2}\sum_{i\in V} k_{ii} x_i^2 + \sum_{i\in V} h_i x_i - \sum_{i\neq j} k_{ij} x_i x_j,$$

where c is appropriate normalizing constant. Compared to the log-linear model, the log-density is additively decomposed of *quadratic main effects* with coefficients $-\frac{1}{2}k_{ii}$, *linear main effects* with coefficients h_i, and *quadratic interactions* with coefficients $-k_{ij}$. Observe that the interaction terms of the highest order involve pairs of variables, and there are no terms involving groups of variables with more than two elements. This is in contrast to the discrete case and it follows in particular that within the normal distribution there are no hierarchical interaction models which are not graphical. So it is an MRF.

Given the interaction graph and a sample (of more than d elements), we want to fit a (Gaussian) distribution so that X_i is conditionally independent of X_j given the remaining variables, denoted by $X_i \perp\!\!\!\perp X_j \,|\, \mathbf{X}_{V\setminus\{i,j\}}$, whenever there is no edge between i and j in G. (Actually, this is the pairwise Markov property, which is equivalent to the local and global Markov properties, as we have a positive distribution.) That is, we want to estimate the mean value parameters ($\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) from the iid. sample $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $(n > d)$, such that the concentration matrix has zero entries in the no-edge positions: $k_{ij} = 0$ whenever $\{i, j\} \notin E$.

This can be done by the covariance selection model: it can be proven (see Theorem 5.3 of [14]) that under this model the ML-estimate of the parameters is: $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$ and that of the restricted covariance matrix $\boldsymbol{\Sigma}^* = (\sigma^*_{ij})$ can be calculated as follows. We estimate the entries in the edge-positions as in the saturated model (no restrictions):

$$\hat{\sigma}^*_{ij} = \frac{1}{n} s_{ij}, \quad \{i, j\} \in E, \tag{18}$$

where $\boldsymbol{S} = (s_{ij}) = \sum_{\ell=1}^{n} (\mathbf{X}_\ell - \bar{\mathbf{X}})(\mathbf{X}_\ell - \bar{\mathbf{X}})^T$. The other entries (in the no-edge positions) of $\boldsymbol{\Sigma}^*$ are free, but satisfy the model conditions: after taking $\boldsymbol{K}^* = (k^*_{ij}) = {\boldsymbol{\Sigma}^*}^{-1}$ with these undetermined entries, we get the same number of equations for them from $k^*_{ij} = 0$ whenever $\{i, j\} \notin E$. To do so, there are numerical algorithms at our disposal, for instance, the IPS (Iterative Proportional Scaling (see [14], p. 134), already discussed in Section 2.5.

The quintessence of the IPS is that it suffices to state Equations (18) for the cliques:

$$\hat{\boldsymbol{\Sigma}}^*_C = \frac{1}{n} \boldsymbol{S}_C, \quad C \in \mathcal{C},$$

where the subscript indicates that we choose the quadratic (and symmetric) $|C| \times |C|$ submatrix of the underlying covariance or empirical covariance matrix that contains only variables in C. Note that instead of the $n > d$ condition $n > c$ would suffice, where c is the cardinality of the largest (maximum) clique.

Then $\boldsymbol{K}^* = {\boldsymbol{\Sigma}^*}^{-1}$ is the fixed point of the equation $T\boldsymbol{K} = \boldsymbol{K}$, where $T = \prod_{i=1}^{k} T_{C_i}$ with $C_1, \ldots, C_k$ being the cliques of the graph and

$$T_{C_i}\boldsymbol{K} = \boldsymbol{K} + [n(\boldsymbol{S}_{C_i})^{-1} - (\boldsymbol{K}^{-1}_{C_i})^{-1}]^V,$$

where, in general, $[\boldsymbol{M}_{\boldsymbol{C}}]^V$ denotes the $d \times d$ matrix containing the entries of the larger matrix $\boldsymbol{M}$ in the $|C| \times |C|$ block corresponding to $\boldsymbol{C}$, and otherwise zeros.

Then starting with an arbitrary $\boldsymbol{K}^{(0)}$ which contains 0 entries exactly in the no-edge positions of G, the iteration

$$\boldsymbol{K}^{(t)} = T\boldsymbol{K}^{(t-1)}, \quad t = 1, 2, \ldots$$

converges to the inverse of the unique ML estimate:

$$\boldsymbol{K}^{(t)} \to \hat{\boldsymbol{K}}^* = (\hat{\boldsymbol{\Sigma}}^*)^{-1}, \quad t \to \infty.$$

Here an infinite iteration is needed, because in general, there is no explicit solution for the ML estimate. However, in the decomposable case there is no need of running the IPS, but explicit estimates can be given as follows. Recall that if the Gaussian graphical model is decomposable (its concentration graph G is decomposable), then the cliques, together with their separators (with multiplicities), form a JT structure. Denote $\mathcal{C}$ the set of the cliques and $\mathcal{S}$ the set of the separators in G.

Then direct density estimates, like (9), are available:

$$f(\mathbf{x}) = \frac{\prod_{j=1}^{k} f(\mathbf{x}_{C_j})}{\prod_{j=2}^{k} f(\mathbf{x}_{S_j})} = \frac{\prod_{C\in\mathcal{C}} f(\mathbf{x}_C)}{\prod_{S\in\mathcal{S}} f(\mathbf{x}_S)^{\nu(S)}}, \quad \mathbf{x} \in \mathbb{R}^d. \tag{19}$$

There are also exact tests in decomposable models (see [14], p. 149).

The ML estimator of $\boldsymbol{K}$ can be calculated based on the product moment estimators applied for subsets of the variables, corresponding to the cliques and separators. First, introduce the simpler form for $\boldsymbol{K}$, see [14]:

$$\boldsymbol{K} = \boldsymbol{\Sigma}^{-1} = \sum_{C\in\mathcal{C}} [\boldsymbol{K}_C]^V - \sum_{S\in\mathcal{S}} [\boldsymbol{K}_S]^V = \sum_{C\in\mathcal{C}} [\boldsymbol{\Sigma}_C^{-1}]^V - \sum_{S\in\mathcal{S}} [\boldsymbol{\Sigma}_S^{-1}]^V,$$

further,

$$|\boldsymbol{\Sigma}| = \frac{\prod_{C\in\mathcal{C}} |\boldsymbol{\Sigma}_C|}{\prod_{S\in\mathcal{S}} |\boldsymbol{\Sigma}_S|}.$$

Let n be the sample size for the underlying d-variate normal distribution, and assume that $n > d$. For the clique $C \in \mathcal{C}$, let $[\boldsymbol{S}_C]^V$ denote n times the empirical covariance matrix corresponding to the variables $\{X_i : i \in C\}$ complemented with zero entries to have a $d\times d$ (symmetric, positive semidefinite) matrix. Likewise, for the separator $S \in \mathcal{S}$, let $[\boldsymbol{S}_S]^V$ denote n times the empirical covariance matrix corresponding to the variables $\{X_i : i \in S\}$ complemented with zero entries to have an $d \times d$ (symmetric, positive semidefinite) matrix. Then the ML estimator of the mean vector is the sample average (as usual), while the ML estimator of the concentration matrix is

$$\hat{\boldsymbol{K}} = n \left\{ \sum_{C\in\mathcal{C}} [\boldsymbol{S}_C^{-1}]^V - \sum_{S\in\mathcal{S}} [\boldsymbol{S}_S^{-1}]^V \right\};$$

further,

$$|\hat{\boldsymbol{K}}| = n^d \cdot \frac{\prod_{S\in\mathcal{S}} |\boldsymbol{S}_S|}{\prod_{C\in\mathcal{C}} |\boldsymbol{S}_C|}.$$

Again, here the structure of $\boldsymbol{K}$ imitates the junction tree structure, through RZP's. Also, decomposable (multiplicative) models provide the Markov property through a chain, and a factorization, resembling (11), also holds:

$$f(\mathbf{x}) = \prod_{i=1}^{k} f(\mathbf{x}_{R_j} | \mathbf{x}_{S_j}) \tag{20}$$

in the RIP ordering of the cliques, residuals, and separators.

By [7, 21], the same can be done for all members of the *exponential* family.

3.4 Directed model, recursive linear regression, and path analysis

Now some causal relations are built in the covariance selection model. Using the estimated inverse covariance matrix, we build a so-called regression graph, and special constellation of the zeros in the concentration matrix, the RZP will give an ordering of the vertices in which causation may happen. Again, it is important that (marginal) independences are indicated by the zero entries of the covariance matrix (more exactly, entries of the sample covariance matrix which do not differ significantly from 0); whereas, conditional independences can be concluded from the (sample) concentration matrix, or can be supplanted in it via covariance selection.

For $2 \leq k \leq d$, consider the following recursive system of linear equations:

$$\begin{aligned} X_1 + a_{12}X_2 + a_{13}X_3 + \cdots + a_{1d}X_d &= \varepsilon_1 \\ X_2 + a_{23}X_3 + \cdots + a_{2d}X_d &= \varepsilon_2 \\ \vdots &= \vdots \\ X_k + \cdots + a_{kd}X_d &= \varepsilon_k, \end{aligned} \tag{21}$$

where $X_1, \ldots, X_k$ are so-called *endogenous*, and $X_{k+1}, \ldots, X_d$ are fixed or so-called *exogenous* (in other wording, *context*) variables, whereas the errors are $\varepsilon_i \sim \mathcal{N}(0, \delta_i)$ for $i = 1, \ldots, k$ and $\mathbb{E}(\varepsilon_i \varepsilon_j) = 0$ for $i \neq j$. So X_i's are also Gaussians with zero expectations, and for $i = 1, \ldots, k$, X_i depends on $X_{i+1}, \ldots, X_d$ linearly, described by equations (21) with the regression coefficients a_{ij}'s which are estimated based on iid measurements $\mathbf{x}_i = (x_{i1}, \ldots, x_{in})^T \in \mathbb{R}^d$, $i = 1, \ldots, n$. Assume that $n > d$ and the sample means (averages of the coordinates of $\mathbf{x}_i$'s) are zeros. Then the $n \times n$ symmetric sample covariance matrix $\boldsymbol{S}$ has entries $s_{ij} = \frac{1}{n}\mathbf{x}_i^T \mathbf{x}_j$. If $n > d$, $\boldsymbol{S}$ is positive definite with probability 1.

When there are no restrictions on a_{ij}'s, the system is called *complete*, and when some of the a_{ij}'s are restricted to be zero, it is *incomplete*. In both cases the MLE's of the parameters can be obtained by applying the method of least squares to each equation separately (where the variables with coefficients restricted to zero do not enter into the regression). The forthcoming theory guarantees that the equations need not be treated separately, but can be solved simultaneously with a convenient decomposition of the concentration matrix (and of the sample concentration matrix) of X_i's. It is also interesting that which pattern of the parameters restricted to zero makes it possible to use a unique method for the parameter estimation. It will turn out that those are the *decomposable*, in other wording, *multiplicative models* which possess this property, and they are strongly related to the decomposable graphs and JT's. For this purpose, let us form a graph with the variables.

We form the directed graph G on d vertices, which correspond to $X_1, \ldots, X_d$. For $i = 1, \ldots, k$ and $j = i+1, \ldots d$, we draw a directed edge $X_j \to X_i$ if $a_{ij} \neq 0$ (X_j is explanatory for X_i) and there is no edge between them if $a_{ij} = 0$. Assume that between the exogenous variables $X_{k+1}, \ldots, X_d$ all edges are present, but those are bidirected and carry no information for the system. This notation was elaborated in path analysis [33], but here we discuss the topic with the simpler notions of [28]. We can as well think of the biderected edges as undirected, and

sometimes we will also forget the direction of the directed edges, as the criteria for decomposability do not use the direction; however the directions somehow dictate the ordering of the cliques in the JT and a perfect numbering of the vertices.

Now we do not regard $X_{k+1}, \ldots, X_d$ as fixed but as rv's and consider $\mathbf{X} = (X_1, \ldots, X_d) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Assume that $\boldsymbol{\Sigma} > 0$ and so, $\boldsymbol{K} = \boldsymbol{\Sigma}^{-1} > 0$. We complete the system (21) with further $d-k$ complete recursive equations to get its matrix form:

$$\boldsymbol{A}\mathbf{X} = \boldsymbol{\varepsilon} \quad \text{with} \quad \boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_d)^T, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Delta}), \tag{22}$$

where $\boldsymbol{A}$ is a $d \times d$ upper triangular matrix with 1's along its main diagonal, otherwise it contains the a_{ij}'s, and $\boldsymbol{\Delta} = \mathrm{diag}(\delta_1, \ldots, \delta_d)$ is $d \times d$ diagonal matrix with positive diagonal entries.

From Equation (22) we get that

$$\mathbb{E}[(\boldsymbol{A}\mathbf{X})(\boldsymbol{A}\mathbf{X})^T] = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T = \boldsymbol{\Delta}.$$

So given the covariance matrix $\boldsymbol{\Sigma} > 0$, we have to find a decomposition

$$\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T = \boldsymbol{\Delta}, \tag{23}$$

where $\boldsymbol{A}$ is upper triangular (with all 1's along its main diagonal) and $\boldsymbol{\Delta}$ is diagonal matrix with positive diagonal entries. If $\boldsymbol{\Sigma} > 0$, then there is a one-to-one correspondence between $\boldsymbol{\Sigma}$ and the pair $(\boldsymbol{A}, \boldsymbol{\Delta})$, which provides the unique solution of Equation (22). The entries of $\boldsymbol{A}$ and $\boldsymbol{\Delta}$ are related to partial regression coefficients and residual variances, see [28] and our forthcoming reasoning; further, we can state the following.

Proposition 6 (Proposition 1 of [28]) *The following are equivalent:*

1. *The system $(\boldsymbol{A}, \boldsymbol{\Delta})$ of recursive linear equations is complete.*
2. *The covariance matrix $\boldsymbol{\Sigma}$ of $\mathbf{X}$ is unrestricted.*

How to get the decomposition of Equation (23)? Since $\boldsymbol{\Sigma}$ and $\boldsymbol{A}$ are invertible matrices, equivalently we have that $\boldsymbol{\Sigma} = \boldsymbol{A}^{-1}\boldsymbol{\Delta}(\boldsymbol{A}^T)^{-1}$ and

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{A}^T\boldsymbol{\Delta}^{-1}\boldsymbol{A}. \tag{24}$$

In fact, we have to perform the *Cholesky decomposition* (in this form called LDL decomposition) of the symmetric, positive definite matrix $\boldsymbol{K} = \boldsymbol{\Sigma}^{-1}$, to obtain the decomposition

$$\boldsymbol{K} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T,$$

where $\boldsymbol{L}$ is lower triangular, and with the convenient choice of the diagonal matrix $\boldsymbol{D}$, we can achieve that its diagonal entries are 1's. Then $\boldsymbol{A} := \boldsymbol{L}^T$ and $\boldsymbol{\Delta} := \boldsymbol{D}^{-1}$.

For example, to get the first row of $\boldsymbol{A}$, that is the first column of $\boldsymbol{L}$, we just divide the first column (row) of $\boldsymbol{K}$ with k_{11}. So $d_{11}^{-1} = k_{11}$ and $a_{11} = 1$, while

$$a_{1j} = \frac{k_{1j}}{k_{11}}, \qquad j = 2, \ldots, d.$$

In view of Theorem 4, a_{1j}'s $(j = 2, \ldots, d)$ are -1 times the partial regression coefficients of X_1 when regressed with $X_2, \ldots, X_d$. The negative sign comes from the equivalent form

$$X_1 = -a_{12}X_2 - a_{13}X_3 - \cdots - a_{1d}X_d + \varepsilon_1$$

of the first equation of (21), which shows that $-a_{ij}$'s are the partial regression coefficients. Further on, with the notation of Definition 4, we get that

$$a_{ij} = -\beta_{ji\cdot\{i+1,\ldots,d\}}, \qquad j \in \{i+1, \ldots, d\}, \quad i = 1, \ldots, d-1 \tag{25}$$

and

$$\delta_i = \mathrm{Var}(X_i|\mathbf{X}_{\{i+1,\ldots,d\}}), \qquad i = 1, \ldots, d \tag{26}$$

is the conditional (residual) variance of X_i conditioned on the variables $\mathbf{X}_{\{i+1,\ldots,d\}}$.

Due to Cochran [4], there is also a recursion for the correlations and the partial regression coefficients, when the variables are standardized:

$$r_{ij} = \sum_{k=i+1}^{d} \beta_{ki\cdot\{i+1,\ldots,d\}}\, r_{kj}, \quad j \in \{i+1, \ldots, d\}. \tag{27}$$

This is the base of the path analysis.

To treat the incomplete cases, we need a definition, already used for discrete variables and undirected models in Proposition 4.

Definition 5 *Let $I \subseteq \tilde{I}$ be a subset of the set $\tilde{I} = \{(i,j) : 1 \le i < j \le d\}$, i.e., the set of the edges of the complete graph over the d-element vertex-set V. We say that I is **reducible** if for each $(i,j) \in I$ and $h = 1, \ldots, i-1$, we have $(h,i) \in I$ or $(h,j) \in I$ or both.*

*We say that the symmetric $p \times p$ matrix $\boldsymbol{M}$ has **zero structure** with respect to I if the upper-diagonal entries of $\boldsymbol{M}$ are zeros exactly in positions $(i,j) \in I$. If I is reducible, we say that $\boldsymbol{M}$ has a **reducible zero pattern (RZP)**.*

Proposition 7 (Proposition 2 of [28]) *Let $\boldsymbol{\Sigma}_{(1,\ldots,k)}$ and $\boldsymbol{A}_{(1,\ldots,k)}$ be the submatrices of $\boldsymbol{\Sigma}$ and $\boldsymbol{A}$, which remain after deleting rows and columns $1, \ldots, k$. $I_\emptyset := \tilde{I}$ and $I_{(1,\ldots,k)}$ is obtained from $\tilde{I}$ by deleting all pairs (i,j) with $i \in \{1, \ldots, k\}$. With this notation, for every reducible $I \subseteq \tilde{I}$ and $k \in \{0, \ldots d-2\}$, the following are equivalent:*

1. *$(\boldsymbol{\Sigma}_{(1,\ldots,k)})^{-1}$ has zero structure with respect to $I_{(1,\ldots,k)}$.*
2. *$\boldsymbol{A}_{(1,\ldots,k)}$ has zero structure with respect to $I_{(1,\ldots,k)}$.*

In the proof, the formulas (2.8), (2.9) of [28] are used, and the fact, that $\boldsymbol{\Sigma} = \boldsymbol{A}^{-1}\boldsymbol{\Delta}{\boldsymbol{A}^{-1}}^T$ implies (by the nature of the Cholesky decomposition) that

$$\boldsymbol{\Sigma}_{(1,\ldots,k)} = (\boldsymbol{A}^{-1})_{(1,\ldots,k)}\boldsymbol{\Delta}_{(1,\ldots,k)}({\boldsymbol{A}^{-1}}^T)_{(1,\ldots,k)}.$$

So to find the $(k+1)$th row of $\boldsymbol{A}$, only the entries of $\boldsymbol{\Sigma}_{(1,\ldots,k)}$ are used.

The next proposition applies to the $k = 0$ case.

Proposition 8 (Proposition 3 of [28]) *For every reducible $I \subseteq \tilde{I}$ and every pair $(\boldsymbol{A}, \boldsymbol{\Delta}), \boldsymbol{\Sigma}^*$ (latter one denoting the restricted covariance matrix), the following are equivalent:*

1. *$\boldsymbol{A}$ has zero structure with respect to I.*
2. *$(\boldsymbol{\Sigma}^*)^{-1}$ has zero structure with respect to I.*

So only in the labeling of the vertices that gives an RZP it is true that the zeros of $(\boldsymbol{\Sigma}^*)^{-1}$ and $\boldsymbol{A}$ coincide.

To find the ML estimate $\widehat{\boldsymbol{\Sigma}^*}$ of $\boldsymbol{\Sigma}^*$, the covariance selection method of Section 3.3 is applicable. Now we will clarify that which class of covariance selection models can be characterized by a reducible zero pattern in the concentrations (entries of $\boldsymbol{\Sigma}^{-1}$). The author of [28] shows (Section 3) that

1. Every incomplete system $(\boldsymbol{A}, \boldsymbol{\Delta})$ with reducible zero pattern can be equivalently described with a decomposable (multiplicative) covariance selection model.
2. Every decomposable (multiplicative) covariance selection model can, after a proper reordering of the variables, be described by an incomplete system $(\boldsymbol{A}, \boldsymbol{\Delta})$ with reducible zero pattern.
3. The decomposition rule can be derived from a given reducible zero pattern.
4. A reducible zero pattern facilitates computation of the ML estimates of the parameters in a covariance selection model, i.e., there are closed forms for the clique concentrations (see Section 3.3) and hence, for the least squares estimates of the corresponding incomplete system.

It is important that in decomposable models the regression coefficients have the same reducible zero pattern as the concentration matrix (see Propositions 5,6,7 of [28]). The proofs use the equivalent statements of decomposability, which comply with our Proposition 4 of Section 2.3. In Sections 4 and 5 of [28], testing hypotheses to find the zero patterns and a practical example are also considered.

Consequently, the condition for a directed graph to be decomposable (has no sink V pattern) corresponds to the condition of its undirected skeleton to be decomposable (have an RZP). If we order the variables according to this RZP, then the recursive regressions give the same estimates for the correlations by the path coefficients as those expected via ML estimation.

The relation to path analysis is also discussed in [28]. If the Gaussian variables are standardized (they are not only mean centered, but have unit variance), then the estimated regression coefficients are the path coefficients, and with them, there are recursions for the correlations r_{ij}'s (they are consequences of similar recursions between the usual and partial correlations, due to Cochran, see [31]). With Equation (27), and denoting by $\hat{a}_{ik}$'s the ML-estimates of the so-called *path coefficients*,

$$r_{ij} = \sum_{k=i+1}^{d} \hat{a}_{ik} \hat{r}_{kj}$$

holds, if we estimate r_{kj}'s in the usual way. When we start from $\mathbf{\Sigma}^*$, and the correlations are estimated by covariance selection, then

$$r^*_{ij} = \sum_{k=i+1}^{d} \hat{a}_{ik}\hat{r}^*_{kj}$$

is the correlation expected from the path diagram. Then ML ratio test can be used to decide whether the estimated correlations ($\hat{r}_{ij}$'s) and the path correlations (r^*_{ij}'s) differ significantly or not. If not, then the path analysis model with existing arrows fits to our data.

4 Composite models

Models with both discrete and continuous (Gaussian) variables are discussed as Conditional Gaussian (CG) models in [14]. Here we rather focus on graphs with both directed and undirected edges.

4.1 Edge matrices and V's

For edge matrices see page 8 of [31]. An edge-matrix $\mathcal{A}$ corresponding to a DAG is upper triangular, contains 1's along its main diagonal, and for $i < j$ its ij entry is 1 if there is a $j \to i$ edge, and 0, otherwise. If the graph also contains undirected edges, the 1's in adjacency positions appear below the diagonal. On page 9 of [31], it is shown that the matrix

$$\mathcal{A}^- = \mathrm{In}[(2\boldsymbol{I} - \mathcal{A})^{-1}]$$

brings in additional dependences and conditional dependences. Here In is an indicator function that assigns 1 to non-zero entries and 0 to the zero ones. The idea is that inversion means a geometric sum, where the powers of the adjacency matrix (edge matrix minus $\boldsymbol{I}$) introduce 1's between ancestral relations up to the order $d - 1$. In this way, additional 1's will appear in the covariance and concentration graphs which means that there are induced edges in them. For example, the existence of a 'sink' $i \to k \leftarrow j$ for $k < i$, $k < j$ when the distinct vertices i and j are not connected with an arrow, induces an $i \sim j$ edge in the concentration graph, as X_i and X_j are conditionally dependent on X_k and so, on all the remaining vertices.

Such operations are called fill-in, 'moralization' in [13, 14]. These procedures also ensure Markov equivalence between two graphs (partially directed and undirected). Two so-called regression graphs (to be introduced next) are Markov equivalent if they define the same independence structure, i.e., the set of independences implied by the graph that goes into the joint distribution of the variables corresponding to the graph's vertices.

4.2 Regression graphs

A regression graph is, in fact, a *chain graph* that contains both directed and undirected edges in the following way. If we keep only the undirected edges, the

graph falls apart into connected components. The components are numbered such that the last ones (with highest indices) correspond to the so-called *context variables* that are given in the context of the experiment. Typically they form the last connected components, and context variables of the same component are connected with undirected edges based on the concentration graph on them. From the context variables arrows show to variables in the lower index boxes (components), which are primary, secondary, etc. responses. From the response variables arrows may show to the response (target) variable(s) which are in lower index boxes (the primary response variables are in the first box, from the left). Between the non-context and non-response variables in the same connected component, there are dashed lines, which indicate dependences on a covariance base. Variables, connected by dashed lines, are also called to be on equal standing; i.e., there is no dashed line between two variables if they are (marginally) independent (the corresponding entry of the covariance matrix of that component, within that box, is 0). Then we can trace the so-called regressions along the arrows. For traceable regressions, see [29, 30], and our examples in Section 4.3.

The simplest version of a regression graph is the so-called *recursive casual model* of Kiiveri et al. [10]. Here the context variables $\mathbf{X}$, called *exogenous*, are in one component (the last one in our labeling, but the first one in the labeling of the authors); whereas the other variables $\mathbf{Y}$, called *endogenous*, as singletons form the other chain components. Here arrows may show from the exogenous variables to endogenous ones, and from endogenous variables arrows show to other endogenous one. As the variables (vertices) connected by directed edges form a DAG, by section 1.1, there is a topological labeling, here called *recursive ordering* of them, so that a $j \to i$ implies $i < j$, in accord with the [30] paper. Note that, on the contrary, $j \to i$ implies $i > j$ in the Kiiveri et al. [10] paper, where a reversed numbering is used. The authors also prove that every causal graph has at least one so-called *extreme* endogenous vertex, such that there is no directed arrow starting from it, i.e., it is the first one in the topological ordering. Indeed, it is the 'youngest' vertex with no children at all, and also a simplicial one in the skeleton of the DAG. Obviously, the exogenous vertices can have only outgoing arrows, and assume that the endogenous ones are labeled as $Y_1, \ldots, Y_d$ in the ordering of [30]. Again, the first some vertices are extreme (the first one is surely that), which are the targets to be predicted.

Theorem 5 (part of Theorem of [10]) *A strictly positive density (pmf or pdf) $p(\mathbf{x}, \mathbf{y})$ corresponding to the casual recursive graph G (here $\mathbf{x}$ belongs to the states of the exogenous, and $\mathbf{y}$ to those of the endogenous variables), the following are equivalent:*

$$\textit{(RCF)} \iff \textit{(GM)} \iff \textit{(LM)},$$

*where **(GM)** and **(LM)** are the global and local Markov properties for recursive causal graphs (they amalgamate the (DG), (UG), and (DL), (UL) properties), and **(RCF)** means the recursive casual factorization as follows. The exogenous variables form an MRF over the undirected part of the graph (for Gaussian variables with positive definite covariance matrix it always holds) and*

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}) \prod_{i=1}^{d} p(y_i | y_{\mathrm{par}(i)}). \tag{28}$$

Now we are interested in the following: what happens if we forget the directions, and consider the underlying graph G as undirected, by just replacing the directed edges with undirected ones. This undirected graph is the *skeleton* of G. The answer is as follows.

Proposition 9 (Corollary of [10]) *Assume that the recursive casual graph G of Theorem 5 has no sink V configuration, see Section 4.1. Then the conditions (RCF), (GM) and (LM) are equivalent to each other and to the undirected Markov property (UM), i.e., the joint distribution is Markov (MRF) over the undirected skeleton of G. Further, if the graph of the exogenous variables is decomposable, then the skeleton of G is also decomposable.*

The authors in [10] also give a variant of the Cholesky decomposition, that triangulates only for the endogenous variables. They prove the following.

Proposition 10 (Lemma 2 of Kiiveri et al. [10]) *The (positive definite) concentration matrix $\boldsymbol{K}$ of the Gaussian system $(\mathbf{Y}, \mathbf{X})$ of endogenous and exogenous random vectors has a unique representation $\boldsymbol{K} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$ with $\boldsymbol{L}$ and $\boldsymbol{D}$ having the form*

$$\boldsymbol{L} = \begin{pmatrix} \boldsymbol{A}^T & \boldsymbol{O} \\ \boldsymbol{B}^T & \boldsymbol{I} \end{pmatrix}, \qquad \boldsymbol{D} = \begin{pmatrix} \boldsymbol{\Delta}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{C}^{-1}, \end{pmatrix}$$

where $\boldsymbol{A}$ is $d \times d$ upper triangular having 1's along its main diagonal, $\boldsymbol{\Delta}$ is $d \times d$ diagonal with positive diagonal entries; whereas the positive definite $\boldsymbol{C}$ and the identity matrix $\boldsymbol{I}$ are of the dimension of $\mathbf{X}$ (say, q).

Then the entries of $\boldsymbol{A} = (a_{ij})$ and $\boldsymbol{\Delta} = \operatorname{diag}(\delta_1, \ldots, \delta_d)$ are determined by the Equations (25) and (26) like

$$a_{ij} = -\beta_{ji \cdot \{i+1,\ldots,d+q\}}, \qquad j \in \{i+1, \ldots, d+q\}, \quad i = 1, \ldots, d-1$$

and

$$\delta_i = \operatorname{Var}(Y_i | \mathbf{Y}_{\{i+1,\ldots,d\}} \mathbf{X}), \qquad i = 1, \ldots, d-1.$$

The $q \times q$ positive definite matrix $\boldsymbol{C}$ is just the covariance matrix of $\mathbf{X}$, while the $q \times d$ matrix $\boldsymbol{B}$ comes by stopping the Cholesky decomposition after the first d columns/rows of $\boldsymbol{K}$ were eliminated.

A *strict ordering* of the vertex set is defined by [10] as any labeling of the exogenous variables together with a topological labeling of the endogenous ones.

Proposition 11 (Proposition 1 of [10]) *A distribution $p(\mathbf{x}, \mathbf{y})$ satisfies the equivalent conditions of Theorem 5 if and only if for all strict orderings of the vertex set the elements of the associated $\boldsymbol{L}, \boldsymbol{D}, \boldsymbol{C}$ in the Cholesky factorization of $\boldsymbol{\Sigma}^{-1}$ satisfy the zero constraints: the zeros in the exogenous part (of $\boldsymbol{C}^{-1}$) correspond to no-edges (covariance selection model), while zeros of $\boldsymbol{L}$ indicate no directed edges from the endogenous variables to another endogenous, or from an exogenous to an endogenous one.*

The authors of [10] also investigate relation to Structural Equation Modelling (SEM) and establish the following.

Proposition 12 *If $\boldsymbol{K}$ is decomposed as in Proposition 10, then $\mathbf{Y}$ and $\mathbf{X}$ satisfy the linear structural equations*

$$\boldsymbol{A}\mathbf{Y} + \boldsymbol{B}\mathbf{X} = \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ and $\mathbf{X}$ are independent Gaussian random vectors with covariance matrices $\boldsymbol{\Delta}$ and $\boldsymbol{C}$. Conversely, if $\mathbf{Y}$ and $\mathbf{X}$ satisfy the above structural equation, further, if $\boldsymbol{\varepsilon}$ and $\mathbf{X}$ are independent with covariance matrices $\boldsymbol{\Delta}$ and $\boldsymbol{C}$, and if $\boldsymbol{A}$ is upper triangular with 1's along its diagonal and $\boldsymbol{\Delta}$ is diagonal, then the matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\Delta}$ combine as in Proposition 10 to give $\boldsymbol{K}$.

Note that $\boldsymbol{A}$ is not necessarily upper triangular, it is that only if the structural equations are recursive. For more general setups see Jöreskog [9].

Some remarks are in order.

- A DAG has a topological ordering ($j \to i$ means $i < j$ for all directed edges). If the DAG does not contain any sink V, then the undirected skeleton is triangulated, so decomposable, and has a perfect numbering of its vertices (see Proposition 4). This perfect numbering is not unique, but it can be the same as a topological ordering of the DAG. Further, in lack of sink V's, any topological ordering of the DAG gives the RZP. So we may call a *DAG decomposable* if it does not contain sink V.

  Consider a DAG which does not contain any sink V together with a topological ordering of its vertices. The undirected skeleton G is triangulated, so decomposable, and has a JT. The cliques of the JT of G in the RIP ordering can be obtained as follows. Let us form the $M_i := \{i\} \cup \mathrm{par}(i) = \mathrm{cl}(i)$ sets. They will be complete subgraphs (because G is triangulated). Delete those that are contained in another one. The remaining M_i's will be the cliques of the JT.

- If the DAG has sink V, then its moral graph is not necessarily triangulated. Also note that even if a DAG has sink V's, its skeleton can be triangulated, so decomposable, and the adjacency matrix has an RZP in a convenient labeling of the vertices. However, this labeling is not topological in the original DAG, where the direction of the edges correspond to real causation given by the real-life problem.

  For example, let the directed adjacency matrix be

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

  Then the upper diagonal part of $\boldsymbol{A}$ does not have the RZP, due to the sink $2 \to 1 \leftarrow 3$. However, the undirected skeleton of this DAG is triangulated, so has a labeling of the vertices in which it has the RZP (for example, in the 2,1,3,4 permutation of the vertices), but this ordering is not topological in the DAG.

  Note that when the DAG has sink V, then to that a triplet $i \to h \leftarrow j$ corresponds with $h < i < j$ and $a_{hi} \neq 0$, $a_{hj} \neq 0$, but $a_{ij} = 0$, see

vertices 1,2,3 in the above $\boldsymbol{A}$, in contrast to the definition of RZP (see Proposition 4).

- Conversely: let us have a decomposable (triangulated) graph G, and a perfect numbering of its vertices. Let us form a directed graph on the same vertex set in the following way: consider a perfect numbering of the vertices and for $i < j$ we draw a $j \to i$ edge whenever $i \sim j$ in G. This results in a DAG. The perfect numbering of the vertices of G also gives a topological ordering of the DAG's vertices.

  For example, let the directed adjacency matrix be

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

  This DAG has no sink V's, and the given topological labeling of the vertices indeed defines the RZP.

- The message of Proposition 9 is that in lack of sink V's, the skeleton graph is Markov equivalent to the original recursive casual G, and it is also decomposable if the undirected part of G is decomposable. So the directed part can fully describe the independence statements if there are no sink V's in it.

  The other message is that if the exogenous part is decomposable, then the exogenous vertices can form the first cliques of a JT, and the others are formed by the endogenous ones, in their reversed topological ordering. So the topological ordering of the DAG gives the perfect ordering of its decomposable skeleton, provided there are no sink V's in it.

- In case of a DAG, the covariance selection means checking for ZPA-s (zero partial associations/correlations) only of the restricted part of the inverse covariance matrix (between a vertex and the higher labeled vertices). These are not necessarily the same as the zeros of the whole covariance matrix above the main diagonal (used for covariance selection in the undirected graph). However, in lack of sink V's, an RZP exists, in which labeling of the vertices, the ZPA in the DAG and in the undericted graph coincide, i.e., the zeros of $\boldsymbol{A}$ are the same as those in the upper-diagonal part of $\boldsymbol{\Sigma}^{-1}$, in view of the Cholesky decomposition. Therefore, these so-called decomposable graphs have the same ZPA, irrespective whether they are directed or undirected.

- Write's rule:

$$r_{ij} = \sum_{k \to i;\, j \to k} \beta_{ki.\{i+1,\ldots,d\}} r_{kj} = \sum_{k \to i;\, j \to k} \beta_{ki|\{i+1,\ldots,d\}\setminus k} r_{kj}$$

  In the $d = 3$ case, if direct effect $X_3 \to X_1$ and indirect effect $X_3 \to X_2 \to X_1$ are both present, then

$$r_{13} = \beta_{13|2} + \beta_{12|3} r_{23}$$

is the sum of the direct and indirect effect. If the direct effect is close to zero, then only the indirect effect $X_3 \to X_2 \to X_1$ is present, and $X_1 \perp\!\!\!\perp X_3 | X_2$ holds, approximately.

On the other hand, the inditect effect (which is a product) can be zero only if either $X_2 \perp\!\!\!\perp X_3$ or $X_1 \perp\!\!\!\perp X_2 | X_3$ holds (no effect reversal).

Going back to the more general case, [30] formulates more general statements about the Markov equivalences of regression graphs. By [32] (page 11), two different graphs are Markov equivalent if they define the same independence structure. Some more notions are also needed. In a regression graph, above a sink V, other types of so-called *collision V's* exist. These are

$$\circ\text{- - -}\circ\text{- - -}\circ, \qquad \circ \to \circ \leftarrow \circ, \qquad \circ\text{- - -}\circ \leftarrow \circ.$$

Further, a *collision path* has as inner nodes exclusively collision nodes (like the middle nodes in the above collision V's), see [30] (page 222).

Theorem 6 (Theorem 1 of [30]) *Two regression graphs are Markov equivalent if and only if they have the same skeleton and the same set of collision Vs, irrespective of the type of edge.*

Note that a directed 'sink' pattern $i \to k \leftarrow j$ and an undirected pattern $i - k - j$ cannot be equivalent. Consequently, the 'sink' pattern should be filled-in (i and j should be connected in the undirected version). Then they won't have the same skeleton, but they can be Markov equivalent.

Theorem 7 (Theorem 2 of [30]) *A regression graph with a chordal graph for the context variables can be oriented to be Markov equivalent to a DAG on the same skeleton if and only if it does not contain any chordless collision path in four nodes.*

Note that only the following three types of chordless collision paths in four nodes exist:

$$\circ\text{- - -}\circ\text{- - -}\circ\text{- - -}\circ, \qquad \circ \to \circ\text{- - -}\circ \leftarrow \circ, \qquad \circ\text{- - -}\circ\text{- - -}\circ \leftarrow \circ$$

Then the authors of [30] (page 241) define an algorithm (Algorithm 1) for labeling the vertices of a regression graph so that to obtain a Markov equivalent DAG, provided it has a chordal concentration graph (for the context variables) and has no chordless collision path on four nodes. Actually, they use the MCS algorithm for the subgraph spanned by the context variables. These will have the higher labels in the reversed RIP ordering. Then directed edges start from higher number components to lower number ones, while within the components the labeling is immaterial. All the collision V's are replaced by sink V's; and when a dashed line in a component is replaced by an arrow, then they label the endpoints such that the arrow is from a higher label to a lower label one if the labels do not already exist. The authors prove (Lemma 1) that their Algorithm 1 generates a DAG that is Markov equivalent to the original regression graph. Then for the DAG, the recursive linear regression of Section 3.4 can be applied, possibly with linearizing formulas.

4.3 Application

Based on the 2014's Egypt Demographic and Health Survey (EDHS 2014), we examined the effect of background characteristics on the ideal number of children a family thinks manageable to have. The research question is: to what extent do age and education level of married couples affect the conceivable ideal number of children, through intermediate variables (wife's age at first marriage, family's wealth index, total number of births a wife had, and use of contraception). The focus is on a selected random sample of 626 urban married women aged 20-49 years. Figure 5 shows the opposite ordering of the variables as they are entered into the model. The joint distribution of them is approximately multivariate Gaussian, and their labeling is based on the expected relationships between the variables based on the literature. The far right hand box includes the relevant context variables. These are the background variables in the model: husband's and wife's education level in years (for both X_9 and X_8, min = 0, max = 28); husband's age (X_7; min = 20, max = 77) and wife's age (X_6; min = 20, max = 49). The next box from the right contains the two intermediate variables, woman's age at the first marriage (X_5; min = 10, max = 40) and the family wealth index (X_4; min = 1, max = 5). Moving to the next box, the secondary responses are represented. These variables are the number of years the woman has been using any contraception method (X_3; min = 0, max = 28) and the total number of births (X_2; min = 0, max = 9). The first box on the left is the primary response variable, the ideal number of children the family thinks to be optimal (X_1, min = 0, max = 11).

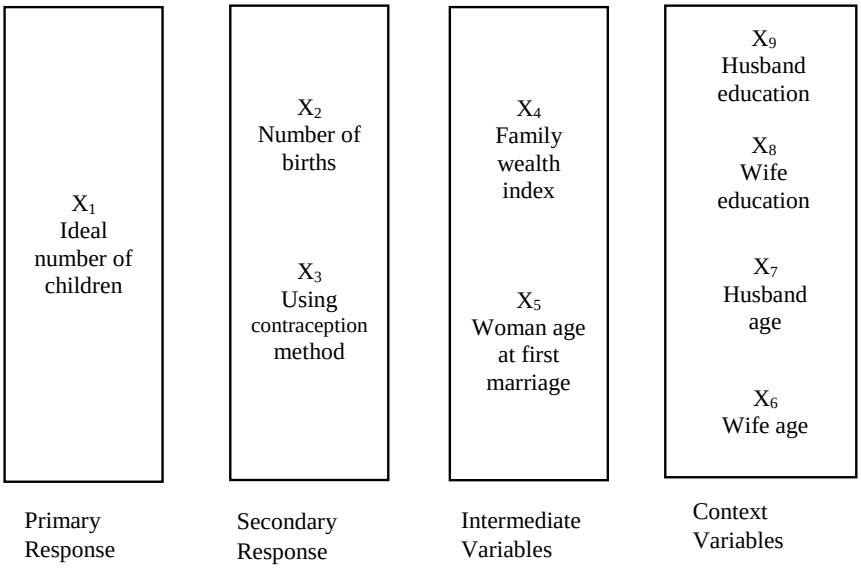

Figure 5: Ordering of the variables as they enter into the model. The variables are selected from the Egyptian ever-married women's questionnaire of EDHS 2014 study.

To examine this model, we first build a regression graph based on the partial

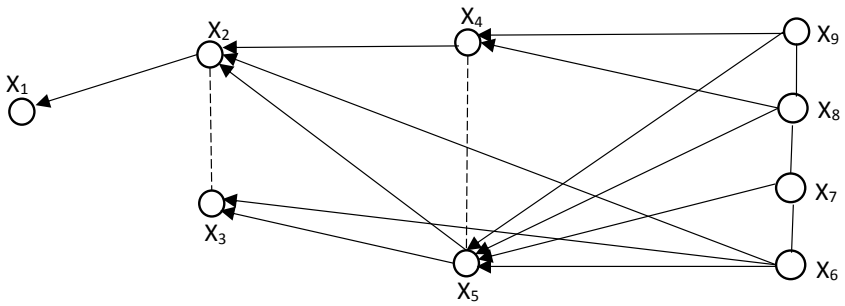

Figure 6: Regression graph of the model. Circles represent the examined continuous random variables.

correlation coefficients between the variables and their ancestors, see Figure 6. The graph summarizes important aspects of the relationships between the variables. It represents the direct causality between some variables by an arrow that goes directly from the explanatory variable pointing to the response. Indirect relationships are represented by a sequence of arrows linking the explanatory variable to an intermediate variable and then continuing to the response variable. As shown in the graph, the context variable components, both age of the married couples, and their education are connected to each other by a solid line that means they are adjacent and influence each other. The intermediate variables in the secondary response components are connected with dashed lines if they are marginally dependent, but are on equal standing.

Later, statistical regression analysis is conducted for the regression graph. The regression results (see Tables 1, 2, 3, 4, 5) confirm the links in the plotted regression graph, due to Corollary 1. The significance of variables in the tested models show which variables are directly explanatory and which are important for generating and predicting a response, and which ones affect only indirectly the response.

The listed variables to the right of a response without an arrow pointing to the response are not essential to improve the prediction of the response when they are used in addition to the directly explanatory variables. As for the conceivable ideal number of children (X_1), only the number of births (X_2) is directly explanatory. The number of births is an important mediator between the woman's current age (X_6), the family wealth index (X_4), her age at first marriage (X_5) and the response (X_1).

The results suggest that the woman's age at first marriage is crucial, and that it is strongly affected by her education. Well educated women are more likely to be older at the first marriage. That reduces the number of births the woman has had, and thus her conceivable ideal number of children. Some of variables are indirectly explanatory. An arrow starts from an explanatory

Explanatory variables	Coeff	S_{coeff}	Std. Error	Sig
Constant	2.427	--	.358	.000
X_2, number of births	--	.432	.044	.000
X_3, using contraception method.	--	-.096	.008	.066
X_4, family wealth index	--	-.018	.051	.661
X_5, woman age at first marriage	--	.056	.013	.236
X_6, wife age	--	-.126	.011	.098
X_7, husband age	--	.045	.008	.512
X_8, wife education	--	-.057	.011	.250
X_9, husband education	--	-.040	.010	.374

R^2 = .38; The model: $X_1 = 2.43 + .43\ X_2$

Table 1: Response X_1, linear regression

Explanatory variables	Coeff	S_{coeff}	Std. Error	Sig
Constant	3.062	--	.318	.000
X_4, family wealth index	--	-.092	.047	.003
X_5, woman age at first marriage	--	-.419	.011	.000
X_6, wife age	--	.628	.010	.000
X_7, husband age	--	-.082	.008	.112
X_8, wife education	--	-.037	.011	.328
X_9, husband education	--	-.053	.010	.132

R^2 = .46; The model: $X_2 = 3.06 - .09\ X_4 - .42X_5 + 0.63\ X_6$

Table 2: Response X_2, linear regression

Explanatory variables	Coeff	S_{coeff}	Std. Error	Sig
Constant	3.97	--	1.678	.018
X_4, family wealth index	--	.030	.251	.435
X_5, woman age at first marriage	--	-.276	.058	.000
X_6, wife age	--	.395	.052	.000
X_7, husband age	--	-.003	.042	.960
X_8, wife education	--	-.022	.055	.642
X_9, husband education	--	-.081	.051	.058

R^2 = .21; The model: $X_3 = 3.97 - .28\ X_5 + .40X_6$

Table 3: Response X_3, linear regression

variable and points, via a sequence of arrows, through intermediate variables, to the response variable. For example, the husband's education (X_9) indirectly affects the conceivable ideal number of children (X_1). It directly affects the family wealth index (X_4), which, in turn, affects directly the number of births

Explanatory variables	Coeff	S_{coeff}	Std. Error	Sig
Constant	3.14	--	.183	.000
X_6, wife age	--	.108	.007	.074
X_7, husband age	--	-.037	.006	.543
X_8, wife education	--	.389	.008	.000
X_9, husband education	--	.075	.008	.005

$R^2 = .19$; The model: $X_4 = 3.14 + .39X_8 + .08\ X_9$

Table 4: Response X_4, linear regression

Explanatory variables	Coeff	S_{coeff}	Std. Error	Sig
Constant	15.58	--	.796	.000
X_6, wife age	--	.561	.032	.000
X_7, husband age	--	-.428	.027	.000
X_8, wife education	--	.240	.034	.000
X_9, husband education	--	.130	.034	.002

$R^2 = .21$; The model: $X_5 = 15.58 + .56\ X_6 - .43X_7 + .24X_8 + .13\ X_9$

Table 5: Response X_5, linear regression

(X_2), while this, in turn, affects the conceivable ideal number of children (X_1).

4.4 Further perspectives

Conditional independences and dependences are captured by regression graphs if the generated distribution shares some properties with a multivariate Gaussian distribution. After a thorough statistical analysis and applying Theorem 7 together with its construction, we construct a DAG. In the topological ordering of the vertices, given by the construction of the theorem, instead of linear, linearized, or logistic regression we take conditional expectation in a nonparametric way, like the ACE (Alternating Conditional Expectation) algorithm [2]. Here we need not alternate, we just take directed conditional expectations, see [8].

Say, we have a data set of cases $\mathbf{x}^{(i)}$ that are multidimensional observations with coordinates labeled as $\mathbf{x}^{(i)} = (x_1^{(i)}, \dots, x_d^{(i)})$, $i = 1, \dots, n$ (we call it corpus). We also have a DAG on d vertices constructed on the above way, based on the corpus. Then a new case $\mathbf{x}^{(n+1)}$ comes with missing variables, we know only the last some coordinates of it. We never know the first coordinate, which is the very target to be predicted, but we know at least the coordinates corresponding to its context variables. So let $1 \leq k < d$ be an integer, so that the last $d - k$ coordinates of our new case are known, and $d - k$ is at least the number of the context variables. Then to predict the first k coordinates, we successively proceed as follows. First,

$$x_k^{(n+1)} := \mathbb{E}(X_k \,|\, \mathbf{x}_{k+1,\dots,d}^{(n+1)}) = \mathbb{E}(X_k \,|\, \mathbf{x}_{\mathrm{par}(k)}^{(n+1)}),$$

where the second equality follows by Markovity. When $k > 1$, we proceed

backward: for $j = k-1, k-2, \ldots, 1$:

$$x_j^{(n+1)} := \mathbb{E}(X_j \,|\, \mathbf{x}_{j+1,\ldots,d}^{(n+1)}) = \mathbb{E}(X_j \,|\, \mathbf{x}_{\mathrm{par}(j)}^{(n+1)}).$$

If our data are from multivariate Gaussian distribution, then the above conditional expectations are linear functions of the variables in the condition, and are obtainable by linear regression (better to say, the coefficients are estimated from the corpus). Otherwise, we take the conditional expectation in a nonparametric way, by the smoothing algorithms discussed in [2]. We illustrate it with a symmetric, bivariate kernel K that depends on some parameters and is also translation invariant. The estimates are as follows.

$$\widehat{x_k^{(n+1)}} = \sum_{i=1}^{n} x_k^{(i)} K(\mathbf{x}_{\mathrm{par}(k)}^{(i)}, \mathbf{x}_{\mathrm{par}(k)}^{(n+1)}) / \sum_{i=1}^{n} K(\mathbf{x}_{\mathrm{par}(k)}^{(i)}, \mathbf{x}_{\mathrm{par}(k)}^{(n+1)})$$

which is a Nadaraya–Watson type local averaging estimate, see [18, 27].

Then for $j = k-1, \ldots, 1$ we continue with

$$\widehat{x_j^{(n+1)}} = \sum_{i=1}^{n} x_j^{(i)} K(\mathbf{x}_{\mathrm{par}(j)}^{(i)}, \mathbf{x}_{\mathrm{par}(j)}^{(n+1)}) / \sum_{i=1}^{n} K(\mathbf{x}_{\mathrm{par}(j)}^{(i)}, \mathbf{x}_{\mathrm{par}(j)}^{(n+1)}). \tag{29}$$

In a greedy way, we could put the already existing estimates for the coordinates $j+1, \ldots, k$ of case $n+1$ into the corpus, and then the summation for i goes from 1 to $n+1$. Likewise, if the next incomplete case $n+2$ comes, we either use the learning sample of n cases, or all of the $n+1$ cases before, etc.

In [2] and [8], other types of smoothings are also introduced, especially for discrete (sometimes categorical) variables. So we could apply smoothings successively for new-coming data through the corpus, and the selection of the kernel should be automated.

If no regression graph is known, but the skeleton is triangulated, we can find a junction tree, and make predictions from separators to residuals according to the factorization

$$p(\mathbf{x}) = \prod_{i=1}^{k} p(\mathbf{x}_{R_j} | \mathbf{x}_{S_j}),$$

where Equation (29) is applied to a multidimensional target.

Consider the ordering of the cliques, obeying the running intersection property with cliques C_j, residuals R_j and separators S_j (indexed from the past to the future), $S_1 = \emptyset$ and $R_1 = C_1$. Assume that we have the coordinates of $\mathbf{x}^{(n+1)}$ corresponding to C_1. Then

$$\mathbf{x}_{R_j}^{(n+1)} := \mathbb{E}(\mathbf{X}_{R_j} \,|\, \mathbf{x}_{S_j}^{(n+1)})$$

for $j = 2, \ldots, k$, where k now denotes the number of cliques. Because of $C_j = R_j \cup S_j$, we so get $\mathbf{x}_{C_j}^{(n+1)}$ and via marginalizing, the new $\mathbf{x}_{S_{j+1}}^{(n+1)}$ is obtained. In a nonparametric way, dropping the new-coming case into the corpus, for $j = 2, \ldots, k$ we have the estimate

$$\widehat{\mathbf{x}_{R_j}^{(n+1)}} = \sum_{i=1}^{n} \mathbf{x}_{R_j}^{(i)} K(\mathbf{x}_{S_j}^{(i)}, \mathbf{x}_{S_j}^{(n+1)}) / \sum_{i=1}^{n} K(\mathbf{x}_{S_j}^{(i)}, \mathbf{x}_{S_j}^{(n+1)}).$$

We plan to make the selection of the best kernel automatic, depending on the type and the range of the variables The above algorithm is also applicable to time series, mainly to Gauss–Markov processes, where the directions of the arrows indicate not only causation but time sequence of the observations. Longitudinal data can also be treated. We also plan to involve SEM and PLS techniques by distinguishing between measurement and latent variables, see e.g., [24].

Acknowledgement

The first author is indebted to Nanny Wermuth for valuable explanations and to András Krámli who first told her about this topic several years ago. The research reported in this paper was supported by the Higher Education Excellence Program of the Ministry of Human Capacities in the frame of Artificial Intelligence research area of Budapest University of Technology (BME FIKP-MI/SC).

References

[1] Bolla, M., *Spectral Clustering and Biclustering. Learning Large Graphs and Contingency Tables*, Wiley (2013).

[2] Breiman, L. and Friedman, J. H., Estimating optimal transformations for multiple regression and correlation, *J. Am. Stat. Assoc.* **80** (1985), 580–619.

[3] Chow, C. K., Liu, C. N., Approximating discrete probability distributions with dependece trees, *IEEE Trans. Inf. Theory* **IT-14** (1968), 452-467.

[4] Cochran, W. G., The omission or addition of an independent variate in multiple linear regression, *J. R. Statist. Soc.*, suppl. **5** (1938), 171-176.

[5] Cox, D. R., Wermuth, N., *Multivariate Dependencies. Models, analysis and interpretation.* Chapman&Hall/CRC (1996).

[6] Darroch, J. N., Lauritzen, S. L., Speed, T. P., Markov Fields and Log-Linear Interaction Models for Contingency tables, *The Annals of Statistics* **8** (3) (1980), 522-539.

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

[8] Győrfi, L., et al., A distribution-free theory of nonparametric regression. Springer Science & Business Media, 2006.

[9] Jöreskog, K. G., Structural equation models in the social sciences. Specification, estimation and testing (1977). In: *Applications of statistics*, ed. P.R. Krishnaiah, North-Holland Publishing Co., 265–287.

[10] Kiiveri , H., Speed, T. P., Carlin, J. B., Recursive casual models, *J. Austral. Math. Soc. (Ser. A)* **36** (1984), 30-52.

[11] Koller, D., Friedman, N., *Probabilistic Graphical Models. Principles and Techniques.* MIT Press (2009).

[12] Lauritzen, S. L., Speed, T. P., Vijayan, K., Decomposable graphs and hypergraphs, *J. Austral. Math. Soc.(Ser A)* **36** (1984), 12-29.

[13] Lauritzen, S. L., Spiegelhalter, D., Local computations with probabilities on graphical structures and their application in expert systems, *J. R. Statist. Soc. B* **50** (1988), 157-224.

[14] Lauritzen, S. L., *Graphical Models*. Oxfor Univ. Press (1995).

[15] Lovász, L., *Combinatorial problems and exercises*. ACM, Akadémiai Kiadó – North Holland, Budapest – Amsterdam (1993).

[16] Lnenicka, R., Matus, F., On Gaussian conditional independence structutes, *Kybernetika* **43** (2007), 323–342.

[17] Móri, F. T., Székely, J. G., Többváltozós statisztikai analízis, Műszaki Könyvkiadó, Budapest (1986) (in Hungarian, Chapter XI. by T. Rudas).

[18] Nadaraya, E. A., On nonparametric estimates of regression functions and regression curves, *Theory of Applied Probability* **10** (1965), 186–190.

[19] Pearl, J., *Causality: Models, Reasoning and Inference*. Cambridge Univ. Press (2000).

[20] Rose, D. J., Tarjan, R. E., Lueker, G. S., Algorithmic aspects of vertex elimination on graphs, *SIAM Journal of Computing* **5** (1976), 266–283.

[21] Sundberg, R., Some results about decomposable (or Markov-type) models for multidimensional contingency tables: distribution of marginals and partitioning of tests, *Scandinavian Journal of Statistics* **2** (1975), 71–79.

[22] Szántai, T., Kovács, E., Application of t-cherry trees in pattern recognition. In: Iantovics B, Radoiu D, Marusteri M, Dehmer M (eds), Broad Research in Artificial Intelligence and Neuroscience (BRAIN), special issue on Complexity in Sciences and Artificial Intelligence, pp. 40-45 (2010).

[23] Tarjan, R. E., Yannakakis, M., Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs and selectively reduce acyclic hypergraphs, *Siam J. Computing* **13** (1984), 566–579.

[24] Tenenhaus, M., Esposito Vinzi, V., Chatelinc, Y-L., Lauro, C., PLS path modeling, *Computational Statistics & Data Analysis* **48** (1), 159–205 (2005).

[25] Wainwright, M. J., Graphical Models and Message-Passing Algorithms: Some Introductory Lectures. In: Mathematical Foundations of Complex Networked Iformation Systems, Lecture Notes in Mathematics 2141, F. Fagnani et al. (eds.), Springer (2015).

[26] M. Wainwright,. M. I. Jordan, Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning 1 (1-2), 1-305 (2008).

[27] Watson, G. S., Smooth regression analysis, *Sankhya* **26** (15) (1964), 359–372.

[28] Wermuth, N., Recursive equations, covariance selection, and path analysis, *J. Amer. Stat. Assoc.* **75** (1980), 963–972.

[29] Wermuth, N., Traceable regressions, *International Statistical Review* (2012) **80** (3), 415–438.

[30] Wermuth, N., Sadeghi, K., Sequences of regressions and their independences, *TEST* **21** (2012), 215-279.

[31] Wermuth, N., Graphical Markov models, unifying results and their interpretation, In: (Balakrishnan, N. et al. eds) Wiley StatsRef: Statistics Reference Online (2015), also ArXiv: 1505.02456.

[32] Wermuth, N., Cox, D. R., Graphical Markov models: Overview, In: (Wright, J., ed.) International Encyclopedia of the Social and Behavioral Sciences, 2nd ed., 10, Elsevier, Oxford, pp. 341–350. (2015), also ArXiv: 1407.7783.

[33] Wright, S., The method of path coefficients, *Ann. Math. Stat.* **5** (3) (1934), 161–215.

Abbreviations

An(A)	ancestral set of the vertex-set A, which is the smallest possible vertex-set (including A) containing all vertices from where a directed path emanates to vertices of A in a directed graph
ant(i)	anteriors of the vertex i in a directed graphs (non-descendants except its parents)
bd(i)	boundary of the vertex i (its neighbors in the undirected, and its parents in the directed case)
BN	Bayesian Network
CG	Conditional Gaussian
cl(i)	closure of the vertex i (it and its biundary)
DAG	Directed Acyclic Graph
DF	Directed Factorization Property
DG	Directed Global Markov Property
DL	Directed Local Markov Property
DP	Directed Pairwise Markov Property
EDHS	Egypt Demographic and Health Survey
iid	independent identically distributed
IPS	Iterative Proportional Scaling
JT	Junction Tree
MCS	Maximal Cardinality Search
ML	Maximum Likelihood
MRF	Markov Random Field
par(i)	parents of the vertex i (from where directed edge shows to it) in a directed graph
pdf	Probability Density Function
pmf	Probability Mass Function
RCF	Recursive Casual Factorization
RIP	Running Intersection Property
rv	random variable
RZP	Reducible Zero Pattern
SEM	Structural Equation Modeling
UF	Undirected Factorization Property
UG	Undirected Global Markov Property
UL	Undirected Local Markov Property
UP	Undirected Pairwise Markov Property