Graphical and log-linear models

Graphical models provide a framework for describing statistical dependencies in (possibly large) collections of random variables. They are multidimensional generalizations of Markov chains. 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. We consider directed and undirected models for discrete, continuous and mixed types of variables, and show how log-linear and Gaussian models can give a general treatment to all of these situations.

1 Directed Graphical Model: Bayesian Network (BN)

BN's are graphical representations of joint distributions. The vertices correspond to random variables (rv's) $X_1, \dots, X_d$, whereas the directed edges to *causal* dependencies between them. The rv's are usually discrete and take 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 (there are 2^d entries in the pmf). However, if we parameterize with the conditional probabilities along the dependencies, we can reduce the calculations, provided the underlying distribution $\mathbb{P}$ is Markov compatible with the directed graph assigned to the rv's in the above way.

We consider a directed, acyclic graph (DAG) G on d vertices with vertex-set $V = \{1, \dots, d\}$. It is important that, in case of a DAG, there is a linear ordering (labeling) of the vertices such that for every directed edge $j \to i$, $j < i$ holds. We use this, so-called (not necessarily unique) *topological labeling* of the vertices.

Let $\mathcal{F}_{\mathrm{Fac}}(G)$ denote the set of all distributions of random vectors $(X_1, \dots, X_d)$ that factorize over G like

$$P(x_1, \dots, x_d) = \prod_{i=1}^{d} P(x_i | x_1, \dots, x_{i-1}) = \prod_{i=1}^{d} P(x_i | x_{\mathrm{pa}(i)}), \qquad (1)$$

where $\mathrm{pa}(i) \subset \{1, \dots, i-1\}$ 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 $x_A = \{x_i : i \in A\}$ and $X_A = \{X_i : i \in A\}$.

On the other hand, let $\mathcal{F}_{\mathrm{Mar}}(G)$ denote the set of all distributions of random vectors $(X_1, \dots, X_d)$ that are Markov with respect to G in the sense that, with the notation $\mu(i) = \{1, \dots, i-1\} \setminus \mathrm{pa}(i)$,

$$X_i \perp X_{\mu(i)} | X_{\mathrm{pa}(i)}, \quad i = 1, \dots, d$$

holds, i.e., X_i (future) and $X_{\mu(i)}$ (past) are independent conditioned on $X_{\mathrm{pa}(i)}$ (present).

This Markov property indicates that every variable is independent of all of its nondescendants (in G), conditioned on its parents. This generalizes the fundamental property of Markov chains (when G is a directed path).

Theorem 1 (Theorem 1 of [13]) *For any DAG G, we have*

$$\mathcal{F}_{Fac}(G) = \mathcal{F}_{Mar}(G).$$

2 Undirected Graphical Model: Markov Random Field (MRF)

MRF's are undirected graph models that explicitly express the conditional independence relationships between the vertices:

$$P(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d) = P(x_i|x_{\mathrm{ne}(i)}),$$

where $\mathrm{ne}(i)$ denotes the set of neighbors (in G) of i.

For an undirected graph G, let $\mathcal{F}_{\mathrm{Mar}}(G)$ denote the set of distributions that are Markov with respect to G in the following symmetric past–future scenario: $X_A \perp X_B | X_S$ holds for any vertex cutset S between disjoint vertex-subsets A and B. In fact, this is the global Markov property, but it coincides with other Markov properties (local) if the underlying distribution is positive (see the Hammersley–Clifford theorem).

Especially, two vertices are conditionally independent if all paths between them are blocked by given vertices. A given vertex can be made conditionally independent of any non-neighboring vertex by observing each neighboring vertex. This is often called the Markov blanket as it "blankets" the vertex from the rest of the graph. There are many possible MRF's for a given $\mathbb{P}$; however, we can draw the tightest MRF using assumptions that each vertex is affected only by its neighbors. More precisely, let $\mathrm{cl}(i) = \{i\} \cup \mathrm{ne}(i)$ denote the closure of vertex i in G. Then, with the above notation, $X_i \perp X_{V \setminus \mathrm{cl}(i)} | X_{\mathrm{ne}(i)}$ holds for any variable X_i.

Now, let $\mathcal{F}_{\mathrm{Fac}}(G)$ denote the set of all distributions that factorize as

$$P(x_1, \ldots, x_d) = \frac{1}{Z} \prod_{C \in \mathbf{C}} \psi_C(x_C) \tag{2}$$

over the undirected graph G, with normalizing constant $Z > 0$ and non-negative *compatibility functions* ψ_C's assigned to the cliques $C \in \mathbf{C}$ of G.

Under *clique* we understand a maximal complete subgraph of G. 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 (usually finite) set $\mathcal{X}_i$. The whole sample space is $\mathcal{X} = \times_{i=1}^{d} \mathcal{X}_i$.

Theorem 2 (Theorem 2 of [13]) ***Hammersley–Clifford theorem.***

$$\mathcal{F}_{Fac}(G) \subseteq \mathcal{F}_{Mar}(G)$$

and equality holds if and only if $P(x_1, \dots, x_d) > 0$, $\forall (x_1, \dots, x_d) \in \mathcal{X}$, i.e., $\mathbb{P}$ has full support.

Note that 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 whenever they are not connected (married). The so obtained *moral graph* is then used in the MRF setup.

We also remark that condition (1) resembles that of (2), since in case of a DAG (2) can be written as

$$\mathbb{P}(x_1, \dots, x_n) = \frac{1}{Z} \prod_{i=1}^{n} f_i(x_i, x_{\mathrm{pa}(i)}) = \frac{1}{Z} \prod_{i=1}^{n} f_i(x_{\mathrm{cl}(i)})$$

where $Z = 1$ and $\mathrm{cl}(i) = \{i\} \cup \mathrm{pa}(i)$ is considered as the closure of vertex i in the DAG.

Gibbs Field (GF)

$\mathbb{P}$ is called a Gibbs distribution if it can be parameterized by a set of positive functions ψ_C's over the cliques of G, by physicists called *clique potentials*, such that for its pmf or pdf the condition (2) holds. By the above Hammersley–Clifford theorem, a GF and MRF are equivalent with regard to the same G, whenever $\mathbb{P}$ is strictly positive.

GF was developed in statistical physics, where the clique potentials are of the form $\psi_C = e^{-f_C}$ with f_C an energy function over values 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. When the cliques are vertices and vertex-pairs (e.g., G is a grid), then the GF gives the classical Ising Model. 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 [6, 7].

3 Log-linear models

3.1 Basic notions

Here the sample space is a contingency table that contains joint observations for (usually not independent) categorical random variables. We shall describe special models in which so-called interactions between the rv's are closely related to their conditional independences and to graphical models.

Let $X_1, \dots, X_d$ be categorical variables, where X_i takes on values in the finite set $\mathcal{X}_i = \{1, \dots, r_i\}$, $i = 1, \dots, d$. The components of the random vector $(X_1, \dots, X_d)$ 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 $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_d$. In fact, $\mathcal{X}$ is a d-dimensional array, the entries of which are d-tuples $x = (x_1, \dots, x_d) \in \mathcal{X}$, 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 $n(x)$, $x \in \mathcal{X}$, where the nonnegative integer $n(x)$ is the number of observations for the random vector $(X_1, \dots, X_d)$ that fall in the cell x out of the total n observations. In other words, n is the sample size, and of course,

$$n = \sum_{x \in \mathcal{X}} n(x).$$

When n is kept fixed, the joint distribution of the counts, $N(x)'s$ as rv's, is multinomial with parameters $p(x)$, $x \in \mathcal{X}$:

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

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

$$\hat{p}(x) = \frac{n(x)}{n}, \quad x \in \mathcal{X}.$$

Now, with some restrictions on the marginal distributions, we shall define more special models. The marginal of the contingency table corresponding to a given subset of the variables $X_\gamma = \{X_i : i \in \gamma\}$, with $\gamma \subset V = \{1, \dots, d\}$, is defined as follows. Let us denote the γ-projection of the d-tuple $x = (x_1, \dots, x_d) \in \mathcal{X}$ by $x_\gamma = \{x_i : i \in \gamma\}$. Then the γ-marginal of the contingency table is given by the marginal counts

$$n(x_\gamma) = \sum_{x' \in \mathcal{X} \colon x'_\gamma = x_\gamma} n(x'), \quad \text{for} \quad x_\gamma \in \mathcal{X}_\gamma = \times_{i \in \gamma} \mathcal{X}_i.$$

So if $|\gamma| = k$, then these counts form a k-dimensional contingency table of $\prod_{i \in \Gamma} r_i$ cells, and there are $\binom{d}{k}$ possible γ-marginals ($k = 1, \dots, d$). Note that the marginal and the conditional distributions are also multinomial whenever the underlying distribution is multinomial. The γ-marginal distribution of the $\{p(x) : x \in \mathcal{X}\}$ distribution is defined by

$$p_\gamma(x_\gamma) = \sum_{x' \in \mathcal{X} \colon x'_\gamma = x_\gamma} p(x'), \quad \text{for} \quad x_\gamma \in \mathcal{X}_\gamma.$$

Likewise, the γ-marginal distribution of any $\{P(x) : x \in \mathcal{X}\}$ distribution (not necessarily multinomial) over the contingency table is denoted by P_γ and defined as

$$P_\gamma(x_\gamma) = \sum_{x' \in \mathcal{X} \colon x'_\gamma = x_\gamma} P(x'), \quad \text{for} \quad x_\gamma \in \mathcal{X}_\gamma.$$

Such distributions can be artificially constructed by fixing certain γ-marginals $\gamma \in \Gamma$, where Γ is a family of subsets of $V = \{1, \dots, d\}$. Denoting the fixed marginal distributions by $\bar{P}_\gamma$ ($\gamma \in \Gamma$), they define a so-called linear family

$$\mathcal{L} = \{P : P_\gamma = \bar{P}_\gamma, \, \gamma \in \Gamma\}$$

(see [3]). In fact, the condition $P_\gamma = \bar{P}_\gamma$ can as well be formulated in terms of functions $f_\gamma : \mathcal{X}_\gamma \to \mathbb{R}$ ($\gamma \in \Gamma$), giving some ANOVA-like conditions for the marginals corresponding to $\gamma \in \Gamma$. See Section 3.3 for examples.

Further, we know that the exponential family that corresponds to this linear family (through any given distribution $\mathbb{Q}$ on $\mathcal{X}$ via its I-projection to the linear family) consists of all distributions that can be represented in product form as

$$P(x) = cQ(x) e^{\sum_{\gamma \in \Gamma} f_\gamma(x_\gamma)} \tag{4}$$

where c is a normalizing constant, and f_γ statistics.

The family of all distributions in the form (4) is called *log-affine family* with interactions in Γ. Frequently $\mathbb{Q}$ is the uniform distribution (i.e., $Q(x) = 1$) that results in the *log-linear model*:

$$\ln P(x) = f_0 + \sum_{\gamma \in \Gamma} f_\gamma(x_\gamma), \tag{5}$$

where the individual terms represent interactions corresponding to $\gamma \in \Gamma$, for they depend on x only through x_γ, and the constant term f_0 corresponds to $\emptyset \in \Gamma$ (it is in accord with the forthcoming hierarchical structure of Γ). This is also in accord with the notation of the Gibbs Field, see Section 2, where $f_C = -f_\gamma$, apart from a constant, if Γ consists of the maximal cliques C's. This is the case in the decomposable models, see the forthcoming Section 3.3.

The representation in (5) is not unique, but it can be made unique if with any $\gamma \in \Gamma$ and $\gamma' \subset \gamma$, the relation $\gamma' \in \Gamma$ also holds. Such log-linear models are called *hierarchical*, and only hierarchical models will be treated in the sequel. 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. The individual values of these functions are usually not the marginal probabilities, however, the γ's in Γ carry important information on the conditional independences of the variables $X_1, \dots, X_d$. For this purpose, we will consider graphs and hypergraphs with vertices assigned to the variables, see Section 3.3.

In hierarchical log-linear models, when Γ is specified with the set

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

there is another, equivalent form of Equation (5) that uses an exponential parametrization and shows that we are in exponential family:

$$P_\theta(x) = \exp\left\{ \sum_{C \in \mathcal{C}} < \theta_C, I_C(x_C) > - Z(\theta) \right\}.$$

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

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

is a vector, and so, θ is a $\sum_{C \in \mathcal{C}} |\mathcal{X}_C|$-dimensional vector, which dimension is usually less than $|\mathcal{X}| = \prod_{i=1}^n |\mathcal{X}_i|$. Further, $Z(\theta)$ is the log-partition function (it does not depend on $x \in \mathcal{X}$), and $< .,. >$ denotes the inner product in the above finite-dimensional spaces. So the canonical statistics $I_C(x_C)$ take on values in $\mathbb{R}^{|\mathcal{X}_C|}$ for every $C \in \mathcal{C}$. 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,

$$I_C = \{I_{C;J}\,,\ J \in \mathcal{X}_C\},$$

where the usual indicator function $I_{C;J}(x_C)$ is 1 if $x_C = J$ and 0, otherwise. Therefore, the $I_{C;J}$'s are canonical statistics, and their sums, i.e., the frequencies $n(J)$'s of the cells within the cliques are the sufficient statistics entering into the parameter estimation. In fact, the ML estimate of the mean parameters is

$$\hat{\mu}_{C;J} = \frac{1}{n} n(J).$$

Hence, these estimates and those of the log-linear probabilities are based on the clique frequencies, and are obtainable by IPS (Iterative Proportional Scaling), see the next section and [13], page 97. In fact, the mean parameters $\mu_{C;J}$'s are the true cell probabilities, expectations of the indicator variables, within the cliques (only those, as many as the θ's).

Exact formulas are also given later for decomposable models.

3.2 ML-estimation in log-linear models

By [3], the ML-estimate of the true distribution obeying the model (3) is also an MI-estimate (minimizing the I-divergence), and it is the I-projection of $\mathbb{Q}$ onto $\mathcal{L}$, which is the only element common to $\mathcal{L}$ and the exponential family (4). From [8] we also know the following:

$$\{N(x_\gamma), \quad x_\gamma \in \mathcal{X}_\gamma, \quad \gamma \in \Gamma\}$$

is a sufficient statistic for the parameters $f_\gamma : \gamma \in \Gamma$ of the log-linear model. Moreover, as we are in exponential family, the γ-marginals of the ML-estimate $\hat{p}$ are equal to their relative frequencies

$$\hat{p}(x_\gamma) = \frac{n(x_\gamma)}{n}, \quad x \in \mathcal{X}, \, \gamma \in \Gamma.$$

For the numerical approximation of the ML-estimate $\hat{p}$ we use the following *Iterative Proportional Scaling* (IPS) algorithm. Note that here instead of the model parameters f_γ's we estimate the cell probabilities under the model's assumptions. We are looking for the estimate in the form

$$\hat{p}(x) = \frac{m(x)}{n}, \qquad x \in \mathcal{X}$$

where m is called expected count or *mean-vector*. Let $m^{(0)}(x)$ be a starting estimate, $x \in \mathcal{X}$. We know that the Γ-marginals of the ML-estimate are equal to the observed marginals in Γ. Therefore, we try to alter $m^{(0)}(x)$ so that its Γ-marginals be equal to the observed marginals in Γ. For this purpose, we fix some order of the γ's in Γ, but it suffices to deal only with the maximal elements of Γ. Indeed, in a hierarchical model, if the γ-marginals of a distribution are equal to the observed ones, then so do the γ'-marginals too, for $\gamma' \subset \gamma$.

The iteration consists of the following cycles: in each cycle each marginal is treated. If there are M maximal interactions in Γ, then in the t-th step of the iteration, where $t = kM + r$, and $0 \le r < M$ counts the inner, while $k = 1, 2, \dots$ the outer cycles, we update the $(r+1)$-th marginal based on the estimate $m^{(t-1)}$:

$$m^{(t)}(x) = m^{(t-1)}(x) \frac{n(x_{\gamma_{r+1}})}{m^{(t-1)}(x_{\gamma_{r+1}})}.$$

Theorem 3 (see [8]) *If $n(x_\gamma) > 0$ and $m^{(0)}(x_\gamma) > 0$, $\forall x \in \mathcal{X}$ and $\forall \gamma \in \Gamma$, then*

$$m^{(t)}(x) \to n\hat{p}(x) \quad as \quad t \to \infty.$$

3.3 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 iteration, 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 can be characterized by the special dependency structure of the variables when we build a graph or hypergraph on them. The so-called decomposable models are strongly related to the MRF (see Section 2), and therefore, the conditional independences between certain subsets of the variables are also encoded in these models.

From now on, our log-lineal model is hierarchical, and we keep only the maximal interactions in Γ. Such a family Γ is called *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. In case of a special structure of the generating class, we can introduce an exact algorithm that goes through the γ's in a definite order. To discuss this, we need some further notions.

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

The *interaction graph* G 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:

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

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

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.

Observe that different connected components of the interaction graph correspond to variables that are mutually 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.*

In other words, the cliques (maximal complete subgraphs) of the interaction graph assigned to the hypergraph with hyperedges as the maximal interactions, give back the maximal interactions. For example, when the generating class is

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

then the interaction graph has the clique $\{1, 2, 3\}$, which is an edge in the clique hypergraph, but not an edge of the hypergraph generated by Γ, so our log-linear model is not a graphical interaction model. However, when the generating class is

$$\Gamma' = \{\{1, 2, 3\}\}, \tag{7}$$

then the interaction graph has the clique $\{1, 2, 3\}$, which is an edge both in the clique hypergraph and in the hypergraph generated by Γ', so our log-linear model is a graphical interaction model.

Theorem 4 (see [8]) *The distribution* $\mathbb{P}$ *obeying the hierarchical log-linear model with generating class* Γ *defines an MRF, i.e.,* $\mathbb{P} \in \mathcal{F}_{Mar}(G)$ *where* G *is the interaction graph corresponding to it, if and only if the log-linear model is graphical.*

Now we investigate special graphical models, the decomposable ones.

Definition 2 *The definition is recursive. The graph G is (weakly) 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 (weakly) decomposable.*

Proposition 1 (Proposition 2.5 of [7]) *The following two properties are equivalent to the fact that G is (weakly) decomposable:*

- *G is **triangulated** (with other words, **chordal**), i.e., every cycle of G with more than 3 vertices has a chord.*
- *G has the following **running intersection property**: we can number the cliques of it to form a so-called **perfect sequence** $C_1, \dots, C_k$ where each combination of subgraphs induced by $H_{j-1} = C_1 \cup \dots \cup C_{j-1}$ and C_j is a decomposition $(j = 2, \dots, k)$, 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 $C_j \setminus S_j = H_j \setminus H_{j-1}$. This sequence of cliques is also called a **junction tree**.*

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

Definition 3 *The hypergraph H is (weakly) decomposable if it is the clique hypergraph of a (weakly) decomposable graph.*

Proposition 2 (see [8], Corollary 7.5) *A log-linear model is a graphical interaction model whenever the hypergraph H, assigned to its generating class Γ, is (weakly) decomposable.*

In this case, the maximal interactions (in Γ) are identical to the cliques in the associated clique hypergraph G.

For example, the model with generating class Γ' of (7) is decomposable, as G is the complete graph, and its clique hypergraph is H. However, the model with generating class Γ of (6) is not decomposable: though G is the complete graph, its clique hypergraph (with the only hyperedge $\{1, 2, 3\}$) is not identical to H (that contains the hyperedges $\{1, 2\}, \{2, 3\}, \{1, 3\}$).

So a sufficient condition for a log-linear model to be *graphical* is that its interaction graph G is (weakly) decomposable (or equivalently, it is *triangulated*), and H is the clique hypergraph of G. In this situation, we can use the following exact (product) estimate for the probabilities, see [7].

Since our interaction graph is (weakly) decomposable, by Proposition 1 we have the perfect sequence $C_1, \dots, C_k$ of the cliques. Then for the true model parameters we have

$$p(x) = \frac{\prod_{j=1}^{k} p(x_{C_j})}{\prod_{j=2}^{k} p(x_{S_j})} = \frac{\prod_{C \in \mathbf{C}} p(x_C)}{\prod_{S \in \mathbf{S}} p(x_S)^{\nu(S)}}, \quad x \in \mathcal{X} \tag{8}$$

where $\mathbf{C}$ is the set of the cliques, $\mathbf{S}$ is the set of the separators, and $\nu(S)$ is the multiplicity of the occurrence of the separator S in the above perfect sequence of the cliques of G.

Hence, the ML-estimate of the mean vector is

$$\hat{m}(x) = \frac{\prod_{j=1}^{k} n(x_{C_j})}{\prod_{j=2}^{k} n(x_{S_j})} = \frac{\prod_{C \in \mathbf{C}} n(x_C)}{\prod_{S \in \mathbf{S}} n(x_S)^{\nu(S)}}, \quad x \in \mathcal{X} \tag{9}$$

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

Hereby we illustrate some particular models via the following examples.

Example 1. Let us consider the rv's X_1, X_2, X_3 taking on values in the finite sets $\mathcal{X}_1 = \{1, \dots, r_1\}$, $\mathcal{X}_2 = \{1, \dots, r_2\}$, $\mathcal{X}_3 = \{1, \dots, r_3\}$. Assume that X_2 and X_3 are independent conditioned on X_1. It means that

$$\mathrm{Prob}(X_2 = j, X_3 = k | X_1 = i) = \mathrm{Prob}(X_2 = j | X_1 = i) \cdot \mathrm{Prob}(X_3 = k | X_1 = i),$$

or with pmf's,

$$\frac{p(i,j,k)}{p(i,*,*)} = \frac{p(i,j,*)}{p(i,*,*)} \cdot \frac{p(i,*,k)}{p(i,*,*)}, \quad i \in \mathcal{X}_1,\ j \in \mathcal{X}_2, k \in \mathcal{X}_3 \tag{10}$$

where $*$ stands for the summation with respect to the other coordinates, thus producing the marginal probability. Here the generating class is

$$\Gamma = \{\{1,2\}, \{1,3\}\}, \tag{11}$$

and the interaction graph has the cliques $\{1,2\}$ and $\{1,3\}$, which are identical to the hyperedges in Γ. So this log-linear model is decomposable with the cliques $\{1,2\}$, $\{1,3\}$, and the only separator $S = \{1\}$ between them. It is also a graphical interaction model, since the hypergraph corresponding to (11) is the clique hypergraph of the interaction graph.

If we simplify Equation (10), we get the formula

$$p(i,j,k) = \frac{p(i,j,*)p(i,*,k)}{p(i,*,*)}, \quad i \in \mathcal{X}_1,\, j \in \mathcal{X}_2, k \in \mathcal{X}_3$$

in accord with (8), and also see the forthcoming Example 4(c).

Consider Example 4.1 of [7], based on an investigation of 237 Danish women performed by the Gallup Institute. X_1 is the childhood experience of physical punishment (yes/no), X_2 is whether they use physical punishment with their children (yes/no), and X_3 is their political affiliation (l,s,r). The model that X_2 and X_3 are independent conditioned on X_1 fits well. Since

$$p(i,j,k) = \alpha(i,j)\beta(i,k),$$

we have that

$$\ln p \in F_{X_1X_2} + F_{X_1X_3},$$

where $F_{X_1X_2}$ is the linear subspace of $2 \times 2 \times 3$ contingency tables with entries depending only on the variables X_1 and X_2 (irrespective of X_3), and $F_{X_1X_3}$ is defined similarly (linear family).

Example 2. If the generating class is

$$\Gamma = \{\{1,3\},\{2,3\},\{3,4\},\{4,5,6\}\}, \tag{12}$$

then the entries of Γ are the cliques of the interaction graph, which are identical to the hyperedges in Γ. So this log-linear model is a graphical interaction model, and it is of course decomposable with the cliques $\{1,3\},\{2,3\},\{3,4\},\{4,5,6\}$, which form a junction tree in this order; the separators are $\{3\},\{3\},\{4\}$, see Figure 1. 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 $x = (x_1,x_2,x_3,x_4,x_5,x_6) \in \mathcal{X}$ and to simplify notation, we did not indicate the missing coordinates ($*$'s) in the marginal probabilities.

Example 3. Let $d = 2$ and for simplicity, denote the cells of the $r_1 \times r_2$ contingency table by (i,j), $i = 1,\dots,r_1$, $j = 1,\dots,r_2$. Then for the cell probabilities, $p(i,j)$'s the log-linear model with generating class $\Gamma = \{\{1,2\}\}$

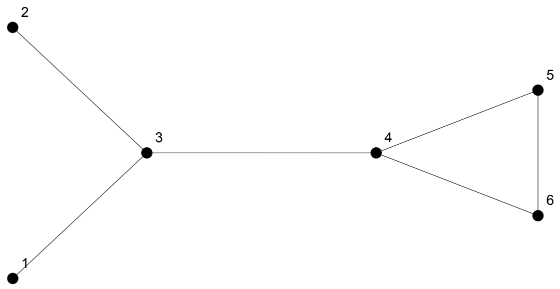

Figure 1: Interaction graph of Example 2, with cliques in (12).

gives the following model equations:

$$\begin{aligned}
f_0 &= \frac{1}{r_1 r_2} \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \log p(i,j) \\
f_1(i) &= \frac{1}{r_2} \sum_{j=1}^{r_2} \log p(i,j) - f_0 \\
f_2(j) &= \frac{1}{r_1} \sum_{i=1}^{r_1} \log p(i,j) - f_0 \\
f_{1,2}(i,j) &= \log p(i,j) - f_1(i) - f_2(j) - f_0.
\end{aligned}$$

It can easily be seen that

$$\log p(i,j) = f_0 + f_1(i) + f_2(j) + f_{1,2}(i,j).$$

The two marginals are independent if and only if $p(i,j)$'s obey the log-linear model

$$\log p(i,j) = f_0 + f_1(i) + f_2(j)$$

with generating class $\Gamma = \{\{1\},\{2\}\}$. However, here the interaction graph is not connected, but consists of two components.

Example 4. Let $d = 3$ and denote the cells of the $r_1 \times r_2 \times r_3$ contingency table by (i,j,k), $i = 1,\dots,r_1$, $j = 1,\dots,r_2$, $k = 1,\dots,r_3$. Then for the cell probabilities, $p(i,j,k)$'s the log-linear model with generating class $\Gamma =$

$\{\{1,2,3\}\}$ gives the following model equations:

$$\begin{aligned}
f_0 &= \frac{1}{r_1 r_2 r_3} \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \sum_{k=1}^{r_3} \log p(i,j,k) \\
f_1(i) &= \frac{1}{r_2 r_3} \sum_{j=1}^{r_2} \sum_{k=1}^{r_3} \log p(i,j,k) - f_0 \\
f_2(j) &= \frac{1}{r_1 r_3} \sum_{i=1}^{r_1} \sum_{k=1}^{r_3} \log p(i,j,k) - f_0 \\
f_3(k) &= \frac{1}{r_1 r_2} \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \log p(i,j,k) - f_0 \\
f_{1,2}(i,j) &= \frac{1}{r_3} \sum_{k=1}^{r_3} \log p(i,j,k) - f_1(i) - f_2(j) - f_0 \\
f_{1,3}(i,k) &= \frac{1}{r_2} \sum_{j=1}^{r_2} \log p(i,j,k) - f_1(i) - f_3(k) - f_0 \\
f_{2,3}(j,k) &= \frac{1}{r_1} \sum_{i=1}^{r_1} \log p(i,j,k) - f_2(j) - f_3(k) - f_0
\end{aligned}$$

$$f_{1,2,3}(i,j,k) = \log p(i,j,k) - f_1(i) - f_2(j) - f_3(k) - f_{1,2}(i,j) - f_{1,3}(i,k) - f_{2,3}(j,k) - f_0.$$

It can easily be seen that

$$\log p(i,j,k) = f_0 + f_1(i) + f_2(j) + f_3(k) + f_{1,2}(i,j) + f_{1,3}(i,k) + f_{2,3}(j,k) + f_{1,2,3}(i,j,k).$$

Then the conditional independences can be planted, via the following generating classes, into the model.

(a) The three 1-dimensional marginals are independent if and only if $p(i,j,k)$'s obey the log-linear model

$$\log p(i,j,k) = f_0 + f_1(i) + f_2(j) + f_3(k)$$

with generating class $\Gamma = \{\{1\},\{2\},\{3\}\}$. Here the interaction graph is again not connected, but consists of three components.

(b) If (a) does not hold, then the 1-dimensional marginal of the first variable is independent of the 2-dimensional marginal of the second and third ones if and only if $p(i,j,k)$'s obey the log-linear model

$$\log p(i,j,k) = f_0 + f_1(i) + f_2(j) + f_3(k) + f_{2,3}(j,k)$$

with generating class $\Gamma = \{\{1\},\{2,3\}\}$. Here the not connected interaction graph consists of two components.

(c) If neither the first and second, nor the first and third marginals are independent, then the second and third marginals are conditionally independent conditioned on the first one if and only if $p(i,j,k)$'s obey the log-linear model

$$\log p(i,j,k) = f_0 + f_1(i) + f_2(j) + f_3(k) + f_{1,2}(i,j) + f_{1,3}(i,k)$$

with generating class $\Gamma = \{\{1,2\},\{1,3\}\}$. Here the interaction graph is connected at last, and the cell probabilities are estimated like in *Example 1.*

Similar statements hold for the permutations of the variables, and we want to emphasize that our primary interest is not to estimate the parameters f_i's, but to estimate the cell probabilities or certain marginal probabilities under the model assumptions. If the model has restrictions (the generating class is not the whole vertex set), then these estimates are not the usual relative frequencies at all.

3.4 Recursive models

So far, the graph G assigned to our log-linear model was an undirected one and we showed how log-linear models are related to the MRF's. When there are casual dependencies between our variables, then we can build a directed graph on them and relate it to the BN's.

Assume that the variables are numbered as $X_1, \dots, X_d$ in such a way that the variable X_j is considered to be a response to variables $X_1, \dots, X_{j-1}$, but explanatory to variables $X_{j+1}, \dots, X_d$. In this case a directed edge $X_j \to X_i$ shows from the explanatory variables to the responses ($j < i$) and parent–child relations can be established as in the BN's. By its formation, this graph is a DAG, and the so-called *recursive graphical model* associated to the above log-linear model with this ordering of the variables offers the following factorization of the likelihood function:

$$\begin{aligned} L(p) &= c \prod_{x \in \mathcal{X}} p(x)^{n(x)} = c \prod_{x \in \mathcal{X}} \left\{ \prod_{i=1}^{d} p(x_i | x_{\mathrm{pa}(i)}) \right\}^{n(x)} \\ &= c \prod_{i=1}^{d} \prod_{x_{\mathrm{cl}(i)} \in \mathcal{X}_{\mathrm{cl}(i)}} p(x_i | x_{\mathrm{pa}(i)})^{n(x_{\mathrm{cl}(i)})} = \prod_{i=1}^{d} L_i(p) \end{aligned}$$

where recall that for $A \subset V$, x_A denotes the A-projection of $x \in \mathcal{X}$, whereas $\mathrm{cl}(i) = \{i\} \cup \mathrm{pa}(i)$ is the closure of vertex i in the DAG as in Section 2.

The likelihood function is factorized as the product of the functions L_i, each being proportional to the likelihood function obtained when sampling the variables in $\mathrm{cl}(i)$ with fixed $\mathrm{pa}(i)$-marginals. The joint likelihood can be maximized by maximizing the factors separately. Each of these factors is in turn proportional to the likelihood function for the saturated model involving the variables in $\mathrm{cl}(i)$ and therefore, the following ML-estimate is derived in [7].

Theorem 5 (Theorem 4.36 of [7]) *The ML-estimate in a recursive graphical model based on a multinomial sample is given as*

$$\hat{p}(x) = \prod_{i=1}^{d} \frac{n(x_{\mathrm{cl}(i)})}{n(x_{\mathrm{pa}(i)})}, \quad x \in \mathcal{X} \tag{13}$$

with the understanding that $n(x_\emptyset) = n$. (This will appear in the denominator whenever a vertex has no parents.)

More generally, in [7] composite models are also discussed, where directed and undirected edges both occur, or the vertices can correspond to continuously distributed rv's such that their distribution conditioned on the discrete rv's is multivariate normal.

Block-recursive models are applicable for *chain-graphs*, where there is a partition of the vertices into $V_1, \dots V_T$ such that the subgraphs have no directed edges, and between the subgraphs arrows show from $V_i \to V_j$ with $(i < j)$. The likelihood function factorizes according to the chain components V_i's. DAG's are also chain graphs, where each vertex is a singleton class.

3.5 Marked graph notation

Sometimes we want to treat uniquely discrete and continuous variables, further, directed and undirected graphs. The vertices of a marked graph belong to both discrete and continuous rv's (V and V') and usually first we label the discrete ones (if directed, then in the topological ordering), then the continuous ones. (Parent and child are defined in the directed, and neighbors, in the undirected case).

Definition 4 (Strongly decomposable graph) *The definition is recursive. The marked graph G is (strongly) decomposable if it is either a complete graph or its vertex-set (V, V') can be partitioned into disjoint vertex-subsets A, B, C such that*

- *C is a complete subset of (V, V');*

- *$C \subseteq V$ or $B \subseteq V'$.*

- *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 (strongly) decomposable.*

Definition 5 (Weakly decomposable graph) *The definition is recursive. The marked graph G is weakly decomposable if it is either it is a complete graph or its vertex-set (V, V') can be partitioned into disjoint vertex-subsets A, B, C such that*

- *C is a complete subset of (V, V');*

- *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 (strongly) decomposable.*

Strong decomposability implies the weak one, and usually the weak suffices, as it guarantees the existence of a junction tree structure by the following proposition. However, in case of mixed models, usually the decomposability is assumed (if it is the strong one, then a special junction tree is needed: the separators are either discrete, or the residuals are continuous).

Proposition 3 (Proposition 2.5 of [7] and Theorem 4 of [13]) *The following properties are equivalent to the fact that G is weakly decomposable:*

- *G is **triangulated** (with other words, **chordal**), i.e., every cycle of G with more than 3 vertices has a chord.*

- *G has the following **running intersection property**: we can number the cliques of it to form a so-called **perfect sequence** $C_1, \dots, C_k$ where each combination of subgraphs induced by $H_{j-1} = C_1 \cup \cdots \cup C_{j-1}$ and C_j is a decomposition $(j = 2, \dots, k)$, 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 $C_j \setminus S_j = H_j \setminus H_{j-1}$. This sequence of cliques is also called a **junction tree**, and equivalently, the graph has a junction tree. That is, S_j is a subset of all cliques on the path between C_{j-1} and C_j in the tree.*

- *The graph is* ***recursively simplicial****. A vertex is simplicial if its neighbors form a complete subgraph. A non-empty graph is recursively simplicial if it contains a simplicial vertex, and when that is removed, any graph that remains is recursively simplicial. (Possible relation to graphs with strongly maximal cliques.)*

Note that the junction tree is indeed a tree with vertices $C_1, \dots, C_k$ and one less edges, that are the separators $S_2, \dots, S_k$. To form it, basically we can start with any clique, but then label the vertices in a so-called perfect order. The last ones are vertices, representing continuous variables.

Proposition 4 (Proposition 2.17 of [7]) *The following properties are equivalent to the fact that G is strongly decomposable:*

- *the vertices of G admit a* ***perfect numbering****;*
- *The cliques of G can be nubered to form a* ***perfect sequence****.*

See also Lemma 14 of [7].

3.6 Undirected Gaussian graphical models

Let $G = (V, E)$ be an undirected graph with vertex set V and edge set E and let $Y = (Y_i)_{i \in V}$ be a multivariate Gaussian random vector. The conditional independences in this case can be easily described as follows. Let $Y \sim \mathcal{N}_d(\mu, \mathbf{C})$ with expectation (vector) μ and positive definite, symmetric $d \times d$ covariance matrix $\mathbf{C}$ ($d = |V|$). Note that a multivariate Gaussian $\mathcal{N}_d(\mu, \mathbf{C})$ distribution with $\mathbf{K} = \mathbf{C}^{-1} = (k_{ij})$ belongs to the expoential family with canonical parameter $(\mathbf{K}\mu, \mathbf{K})$. The also positive definite, symmetric matrix $\mathbf{K} = \mathbf{C}^{-1}$ is called *concentration matrix*, and its zero entries indicate conditional independences between two components of Y, conditioned on the remaining components.

Proposition 5 *For the above Gaussian random vector: if $i \neq j$, then*

$$Y_i \perp Y_j \,|\, Y_{V \setminus \{i,j\}} \Leftrightarrow k_{ij} = 0.$$

Moreover, $-\frac{k_{ij}}{\sqrt{k_{ii}k_{jj}}}$ is the partial correlation coefficient between Y_i and Y_j after eliminating the effect of the remaining components of Y.

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

$$\ln f(y) = c - \frac{1}{2} \sum_{i \in V} k_{ii} y_i^2 + \sum_{i \in V} h_i y_i - \sum_{i \neq j} k_{ij} y_i y_j,$$

where c is appropriate constant. Compared to the discrete case, 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. The hyperedges are usual edges.

The graphical Gaussian model represented by G is the set of Gaussian distributions for which the maximal interactions are pairwise, and the non-zero k_{ij}'s correspond to $\{i,j\} \in E$ pairs. 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.

Given the interaction graph and a sample (of more than d elements), we want to fit a (Gaussian) distribution so that Y_i is conditionally independent of Y_j given the remaining variables, denoted by $Y_i \perp Y_j \,|\, Y_{V \setminus \{i,j\}}$, whenever there is no edge between i and j in G. That is, we want to estimate the parameters (μ and $\mathbf{C}$) from the i.i.d. sample $X_1, \dots, X_n \sim \mathcal{N}_d(\mu, \mathbf{C})$ $(n > d)$, such that the concentration matrix has zero entries in the no-edge positions: k_{ij} be 0 whenever $\{i,j\} \notin E$. This can be done by the covariance selection model: it can be proven (see Theorem 5.3 of [7]) that under this model the ML-estimate of the parameters is: $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and that of $\mathbf{C} = (c_{ij})$ can be calculated as follows. We estimate the entries in the edge-positions as in the saturated model (no restrictions):

$$\hat{c}_{ij} = \frac{1}{n} s_{ij}, \quad \{i,j\} \in E,$$

where $\mathbf{S} = (s_{ij}) = \sum_{\ell=1}^n (X_\ell - \bar{X})(X_\ell - \bar{X})^T$. The other entries (in the no-edge positions) are free, but after taking $\mathbf{K} = (k_{ij}) = \mathbf{C}^{-1}$ with these undetermined entries, we get the same number of equations for them from $k_{ij} = 0$ whenever $\{i,j\} \notin E$. There are numerical algorithms at our disposal, for instance, the iterative proportional scaling (see [7], p. 134). Actually, the equations can be stated for the cliques, and instead of the $n > d$ condition $n > c$ would suffice, where c is the cardinality of the largest (maximum) clique.

If the graph G is decomposable, and we have a junction tree, then direct estimates, like (8), are available:

$$f(y) = \frac{\prod_{j=1}^k f(y_{C_j})}{\prod_{j=2}^k f(y_{S_j})} = \frac{\prod_{C \in \mathbf{C}} f(y_C)}{\prod_{S \in \mathbf{S}} f(y_S)^{\nu(S)}}, \quad y \in \mathbb{R}^d. \tag{14}$$

There are also exact tests in decomposable models (see [7], p. 149).

3.7 Mixed models

Here the variables can be either quantitative (discrete/ordinal/nominal) or qualitative (continuous/scaled) rv's, and they can capture covariation between the discrete and continuous ones. If they are all continuous, we assume the model of Section 3.6, where standard methods of multivariate statistical analysis are applicable.

In the mixed case, we assume that the conditional distribution of the continuous rv's, conditioned on the discrete ones, is multivariate normal. Hierarchical mixed interaction models are full and regular exponential models, being defined through linear restrictions on the canonical parameters of the saturated model. The ML-estimate of the parameters exists. If the graph is decomposable and undirected, the ML-estimate can be calculated explicitly.

So we have a so-called *marked graph* with vertices corresponding to discrete or continuous variables, and the edges are relations between them. Let V and V' denote the vertices corresponding to the discrete and continuous variables, $|V| := d$, $|V'| := d'$. The observations for the discrete rv's are cell counts of a d-dimensional contingency table $\mathcal{X}$, and we shall denote the cells by x as in Section 3.1. The continuous rv's have values $y \in \mathbb{R}^{d'}$. So the sample space is $\mathcal{X} \times \mathbb{R}^{d'}$.

The joint distribution of the $d + d'$ variables is given with the density f, the natural logarithm of which, at (x, y), is the following.

$$\ln f(x,y) = g(x) + h(x)^T y - \frac{1}{2} y^T K(x) y$$

where $x \in \mathbb{R}^d$, $y \in \mathbb{R}^{d'}$, $g(x) \in \mathbb{R}$, $h(x) \in \mathbb{R}^{d'}$ $(\forall x \in \mathcal{X})$, ^{T} denotes the transposition, and $K(x)$ is a $d' \times d'$, positive definite matrix $(\forall x \in \mathcal{X})$. Actually, $K(x)$ is the inverse covariance matrix, called concentration matrix, of the multivariate Gaussian random vector of the continuous rv's occurring together with the joint value x of the discrete ones.

Such a distribution is called *CG distribution*, and it is named homogeneous if $K(x) = K$ $(\forall x \in \mathcal{X})$. The marginal of a CG distribution is not always a CG distribution, but the conditional distribution of any subset of the continuous rv's conditioned on any discrete one is multivariate Gaussian. More precisely, the conditional distribution of the continuous rv's, conditioned on that the discrete ones taking on value in cell x, is $\mathcal{N}_{d'}(K(x)^{-1}h(x), K(x)^{-1})$. We estimate the covariance matrix $K(x)^{-1}$ as the empirical covariance matrix, based on the y_i part of the sample entries (x, y_i)'s.

In graphical interaction models, since the CG distributions are strictly positive, the different Markov properties are all equivalent (we use the global one).

By Proposition 4, we have the following factorization of the joint density into weak marginals to cliques and separators:

$$f(x) = \prod_{j=1}^{k} \frac{f_{C_j}(x_{C_j})}{f_{S_j}(x_{S_j})}.$$

Within each of these complete sets we can further partition the variables into discrete and continuous to obtain

$$f(x,y) = \prod_{j=1}^{k} \frac{p(x_{C_j})}{p(x_{S_j})} \prod_{j=1}^{k} \frac{f_{C_j}(y_{C_j}|x_{C_j})}{f_{S_j}(y_{S_j}|x_{S_j})},$$

where $S_1 = \emptyset$ and $f_\emptyset = 1$.

With special interactions, we get an interaction graph G on the vertex set $V \cup V'$ (in the homogeneous model, the continuous rv's are fully connected if K is positive definite). If G is decomposable, then with the perfect sequence $C_1, \dots, C_k$ of the cliques and the separators S_j's, we have the following estimates for the cell probabilities and the conditional densities, provided that for every clique C: $n(x_C) > 0$ and the empirical covariance matrix, based on the y_i part of the sample entries (x_C, y_i) is positive definite (it holds almost surely whenever $n(x_C) > |C \cap V'|$). Therefore, by Proposition 6.21 of [7], the ML estimates are as follows.

$$\hat{p}(x) = \prod_{j=1}^{k} \frac{n(x_{C_j \cap V})}{n(x_{S_j \cap V})}$$

with the understanding that $n(x_\emptyset) = n$; further,

$$\hat{f}(y|x) = \prod_{j=1}^{k} \frac{\hat{f}_{x_{C_j \cap V}}(y_{C_j \cap V'}|x_{C_j \cap V})}{\hat{f}_{x_{S_j \cap V}}(y_{S_j \cap V'}|x_{S_j \cap V})}.$$

Based on these, we can estimate $h(x)$ and $K(x)$ for $x \in \mathcal{X}$.

Note that here the running intersection $C_1, \dots, C_k$ exhausts the cliques, but for the separators and the remaining parts there are additional requirements (they should contain all discrete or continuous vertices), see [7].

Note that there are exact tests in decomposable models, and CG regressions.

4 Mode prediction

Let $X_1, \dots, X_d$ 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 $x_2, \dots, x_d$ of the others. If x_{1i} denotes the ith possible value of X_1, we are looking for the conditional probabilities

$$p(x_{1i}|x_2, \dots, x_d) = \frac{p(x_{1i}, x_2, \dots, x_d)}{p(x_2, \dots, x_d)}, \quad i = 1, \dots, r_1 \tag{15}$$

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 $X_2, \dots, X_d$ are possible symptoms, and X_1 is the diagnosis, then x_{1i^*} is the most likely diagnosis under the given symptoms. In the child support example, if $X_2, \dots, X_d$ describe the status of the father (salary category, number of children, years spent together, remarried:yes/no), X_1 is the most likely category of the child support fee he has to pay.

Input of the algorithm

- the observed *frequency counts* $n(x_1, \dots, x_d)$: there could be $r_1 \cdots r_d$ frequencies, but we keep only the positive ones. The sum of them is n.
- The graphical interaction structure: a graph with vertices $X_1, \dots, X_d$ that are organized into a *junction tree* structure: the vertices of this junction tree are the cliques C (maximal complete subgraphs) and the edges are the separators S between them (the number of separators is one less than the number of cliques, and they can occur with multiplicities). Such a structure is guaranteed if the graph (with edges as interactions) is triangulated (chordal), but with $d \le 30$ we can build it manually.

Output of the algorithm: The conditional probabilities (15) and the mode of X_1. As the denominator does not depend on i, we only consider the numerator:

$$p(x_{1i}, x_2, \dots, x_d) = p_i(\mathbf{x}) \propto \frac{\prod_C n(\mathbf{x}_C)}{\prod_S n(\mathbf{x}_S)}.$$

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,

$$p(x_{1i}, x_2, \dots, x_d) = p_i(\mathbf{x}) \propto q_i := \frac{\prod_{C:\, X_1 \in C} n(\mathbf{x}_C)}{\prod_{S:\, X_1 \in S} n(\mathbf{x}_S)}.$$

Eventually,

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

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

Note that the estimate 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.

5 Algorithms

Exact algorithms are used for estimating marginals, modes, and likelihoods, e.g., message-passing, sum-product, and max-product algorithms for trees and factor graphs (bipartite graphs to facilitate the description of the clique memberships of the vertices). Graphical models are widely used in statistical machine learning and artificial intelligence; further, in statistical physics, social sciences, communication, information, and network control theory.

The Belief Propagation Algorithm [5, 11] is capable to estimate marginal conditional probabilities based on some evidences (given the values of some X_i's) in BN's. If there are no evidences, it estimates the marginals. The algorithm is an iteration, in due course of which the processors (vertices) send informations along the edges of the so-called factor graph, to be constructed for this convenience. The Belief Propagation Algorithm converges for DAG's, and it is a special case of the Sum-Product Algorithm, for which advanced versions were developed, e.g., the Turbo Codes [1], and relations to information theoretical coding and the Shannon entropy were recovered.

These algorithms are as well applicable to directed and undirected graphs, through moralization. In [7], Markov Chain Monte Carlo (MCMC) methods are also introduced to find the mode estimates, i.e., the most probable category-configurations, provided our distribution is positive. If not, other possibilities are available (they treat the variables in blocks), see programs like BUGS, see [13]. Missing data can be treated via the EM-algorithm.

6 Conclusions

We saw that the graphical hierarchical log-linear models are identical to $\mathcal{F}_{\mathrm{Mar}}(G)$, where G is the interaction graph corresponding to the generating class Γ of the model. On the one hand, under the conditions of the Hammersley–Clifford theorem the (positive) joint distribution factorizes according to the cliques of G and in this case, $\mathcal{F}_{\mathrm{Fac}}(G) = \mathcal{F}_{\mathrm{Mar}}(G)$. On the other hand, when our graphical interaction model is especially decomposable, a factorization over the cliques and separators is possible, no matter whether all cell counts are positive or there are zeros among them. Recall, that G is (weakly) decomposable if and only if it is chordal. How can a non-chordal graph be made chordal with adding the fewest possible edges, there are numerical algorithms at our disposal, e.g., [12].

When the number of variables (d) is not too large, it is not hard to find out whether G is decomposable. If not, we may triangulate it, if yes, we can find a perfect sequence of cliques in it. Since we are interested only in some special cell probabilities, the number of the variable categories can be large. When we have categorical variables with finite state space, the most general are the estimates of (9) in the directed, and of (13) in the undirected case.

When d is large, to find the cliques, though this problem is NP-complete, numerical approximation algorithms are available. In the computer science literature, under clique a complete subgraph is understood and what is called clique by graph theorists, they call it *maximal clique.* This so-called maximal clique problem is widely treated in computer science literature, starting from the seminal paper [9]. For example, the replicator dynamics algorithm of [10] can find dominant sets (which are generalizations of maximal cliques for edge-weighted graphs) with an iteration. Under some conditions, the iteration converges to the characteristic vector of a maximal clique (in the unweighted case), depending on the starting. There can be many overlapping maximal cliques in a graph, and the one (it is also not necessarily unique) with maximal cardinality is called maximum clique by theoretical computer scientists.

In the mixed case, to estimate the densities classical methods of Gaussian based multivariate statistics, nonparametric methods, or the ACE algorithm of [2] can be used.

References

[1] Berrou, C., Glavieux, A., Thimajshima, P., Near Shannonlimit Error Correcting Coding and Decoding: Turbo Codes. In Proc. IEEE Int. Conf. Communications, Geneva, Switzerland (1993), pp. 1064-1070.

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

[3] Csiszár, I., Shields, P., Information Theory and Statistics: A Tutorial, In: Foundations and Trends in Communications and Information Theory, Vol. 1 Issue 4 (2004), Now Publishers, USA.

[4] Gehrmann, H. and Lauritzen, S. L., Estimation of means in graphical Gaussian models with symmetries, *Ann. Stat.* **40** (2012), 1061-1073.

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

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

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

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

[9] Motzkin, T. S., Straus, E. G., Maxima for graphs and a new proof of a theorem of Turán, *Canad. J. Math.* **17** (1965), 533–540.

[10] Pavan, M., Pelillo, M., Dominant sets and pairwise clustering, *IEEE Trans. Pattern Anal. Machine Intell.* **29**(1) (2007), 167–172.

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

[12] 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.

[13] 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).