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) $X_1, \ldots, X_d$, 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 $V = \{1, \ldots, d\}$. 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 $(X_1, \ldots, X_d)$ that factorize over G like

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

where $\mathrm{pa}(i) \subset \{1, \ldots, i-1\}$ 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 $(X_1, \ldots, X_d)$ that are Markov with respect to G in the sense that, with the notation $\mu(i) = \{1, \ldots, i-1\} \setminus \mathrm{pa}(i)$,

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

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.

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

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

over the undirected graph G, with normalizing constant $Z > 0$ and non-negative *compatibility functions* ψ_C's assigned to the cliques $C \in \mathcal{C}$ 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 $P(x_1, \ldots, x_d) > 0$, $\forall (x_1, \ldots, x_d) \in \mathcal{X}$, *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

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

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 $X_1, \ldots, X_d$ 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 $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$. In fact, $\mathcal{X}$ is a d-dimensional array, the entries of which are d-tuples $x = (x_1, \ldots, x_d) \in \mathcal{X}$, 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 $n(x)$, $x \in \mathcal{X}$, where the nonnegative integer $n(x)$ is the number of observations for the random vector $(X_1, \ldots, X_d)$ 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 expected count $m(x)$ 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 $\gamma \subset V = \{1, \ldots, d\}$, is defined as follows. Let us denote the γ-projection of the d-tuple $x = (x_1, \ldots, x_d) \in \mathcal{X}$ by $x_\gamma = \{x_i : i \in \gamma\}$. Then the γ-marginal of the contingency table is given by the marginal counts

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

So 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 if $|\gamma| = k$ $(k = 1, \ldots, d)$. Note that

the marginal and the conditional distributions are also multinomial. The γ-marginal distribution of the $\{p(x) : x \in \mathcal{X}\}$ distribution is defined by

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

By information theoretical tools it can be proven that the only model that accommodates restrictions for the margins is the *log-linear model*:

$$\ln P(x) = f_0 + \sum_{\gamma \in \Gamma} f_\gamma(x_\gamma), \tag{4}$$

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).

The representation in (4) 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 $X_1, \ldots, X_d$. For this purpose, we will consider graphs and hypergraphs with vertices assigned to the variables, see Section 3.2.

3.2 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 we can give the estimate of the cell probabilities under the model's assumptions 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 $V = \{1, \ldots, d\}$, 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

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

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

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

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 3 (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 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** $C_1, \ldots, C_k$ where each combination of subgraphs induced by $H_{j-1} = C_1 \cup \cdots \cup C_{j-1}$ and C_j is a decomposition $(j = 2, \ldots, k)$, 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 $C_j \setminus S_j = H_j \setminus H_{j-1}$. This sequence of cliques is also called a **junction tree**.*

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.*

For example, the model with generating class Γ' of (6) is decomposable, as G is the complete graph, and its clique hypergraph is H. However, the model with generating class Γ of (5) is not decomposable: though G is the complete graph, its clique hypergraph (with the only hyperedge $\{1,2,3\}$) is not identical to H (containing 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

$$p(x)=\frac{\prod_{j=1}^k p(x_{C_j})}{\prod_{j=2}^k p(x_{S_j})}=\frac{\prod_{C\in\mathcal{C}} p(x_C)}{\prod_{S\in\mathcal{S}} p(x_S)^{\nu(S)}},\quad x\in\mathcal{X} \tag{7}$$

where $\mathcal{C}$ is the set of the cliques, $\mathcal{S}$ 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

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

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,

$$\frac{p(i,j,k)}{p(i,*,*)}=\frac{p(i,j,*)}{p(i,*,*)}\cdot\frac{p(i,*,k)}{p(i,*,*)},\quad i\in\mathcal{X}_1,\ j\in\mathcal{X}_2,k\in\mathcal{X}_3 \tag{9}$$

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

$$\Gamma=\{\{1,2\},\{1,3\}\}, \tag{10}$$

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 (10) is the clique hypergraph of the interaction graph.

If we simplify Equation (9), 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 (7), and also see the forthcoming Example 4(c).

Example 2. If the generating class is

$$\Gamma = \{\{1,3\},\{2,3\},\{3,4\},\{4,5,6\}\}, \tag{11}$$

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.

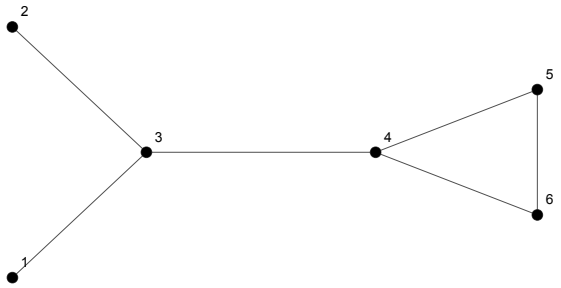

Figure 1: Interaction graph of Example 2, with cliques in (11).

Example 3. Let $d = 2$ and for simplicity, denote the cells of the $r_1 \times r_2$ contingency table by (i,j), $i = 1,\ldots,r_1$, $j = 1,\ldots,r_2$. Then for the cell

probabilities, $p(i,j)$'s the log-linear model with generating class $\Gamma = \{\{1,2\}\}$ 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), $i = 1,\ldots,r_1$, $j = 1,\ldots,r_2$, $k = 1,\ldots,r_3$. 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:

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

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.

3.3 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 $X_1, \ldots, X_d$ in such a way that the variable X_j is considered to be a response to variables $X_1, \ldots, X_{j-1}$, but explanatory to variables $X_{j+1}, \ldots, X_d$. 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 4 (Theorem 4.36 of [7]) *The ML-estimate in a recursive graphical model based on a multinomial sample is given as*

$$\hat{p}(x) = \prod_{i=1}^{d} \frac{n(x_{\mathrm{cl}(i)})}{n(x_{\mathrm{pa}(i)})}, \quad x \in \mathcal{X} \tag{12}$$

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 $V_1, \ldots V_T$ 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.

3.4 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 graphical Gaussian model represented by G is the set of Gaussian distributions for which the maximal interactions are pairwise, as can be seen from formula (5.13) of [7]. 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 $|V|$ 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 direct edge between i and j in G, but any path between them goes through the remaining vertices. This can be done by the covariance selection models, via ML and iterative proportional scaling. If the graph G

is decomposable, and we have a junction tree, then direct estimates, like (7), are available:

$$f(y) = \frac{\prod_{j=1}^{k} f(y_{C_j})}{\prod_{j=2}^{k} f(y_{S_j})} = \frac{\prod_{C\in\mathcal{C}} f(y_C)}{\prod_{S\in\mathcal{S}} f(y_S)^{\nu(S)}}, \quad y \in \mathbb{R}^{|V|}. \tag{13}$$

Note that a multivariate Gaussian $\mathcal{N}_{|V|}(\mu, \Sigma)$ distribution with concentration matrix $\mathbf{K} = \Sigma^{-1} = (k_{ij})$ belongs to the expoential family with canonical parameter $(\mathbf{K}\mu, \mathbf{K})$. It holds that $Y_i \perp Y_j \,|\, Y_{V\setminus\{i,j\}}$ if and only if $k_{ij} = 0$, see [4].

3.5 Mixed models

Here the variables can be either quantitative (discrete) or qualitative (continuous) rv's, and they can capture covariation between the discrete and continuous ones. If they are all continuous, we assume the model of Section 3.4, 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.

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 $C_1, \ldots, C_k$ 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,

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

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 $C_1, \ldots, C_k$ 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].

4 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.

5 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 (8) in the directed, and of (12) 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).