# Chapter 8

# Learning

This chapter is devoted to the methods of learning CI structures on the basis of data. However, it is not the aim of this text to make an overview of numerous methods of learning graphical models from the literature. Rather, the purpose of this chapter is to show how to apply the method of structural imsets to learning DAG models and to indicate that this approach can be extended to learning more general classes of models of CI structure.

Some thoughts about the learning methods which are based on statistical CI tests are mentioned in Section 8.1. The next two sections of the chapter contain an analysis of certain DAG-model learning algorithms based on the maximization of a quality criterion, and also deal with related questions. It is argued in Section 8.4 that the method of structural imsets can be applied in this area as well.

## 8.1 Two approaches to learning

There is plenty of literature about learning graphical models of CI structure - both in the area of statistics and in the area of artificial intelligence. For an overview of some of these methods see § 1.4 and § 4 of [13] and § 3 of [19]. The literature concerns learning UG models, DAG models and DAG models with hidden variables; most attention is devoted to learning DAG models.

In my view, the algorithms for learning graphical models can be divided into two groups on the basis of the fundamental methodological approach.

- Some of these algorithms are based on *significance tests* between two statistical models (of graphical CI structure). These significance tests often correspond to statistical tests for the validity of some CI statements.

- Other algorithms are based on the maximization of a suitable *quality criterion* designed by a statistician. On the basis of the way to derive the criterion, algorithms of this kind could be further divided into those based on the Bayesian approach and those which stem from a frequentist point of view.

Nevertheless, some algorithms may be included in both groups because they can be interpreted in both ways and there is a simulation method applicable to learning graphical models which does not belong to either of these two groups.

**Data faithfulness assumption**

Typical examples of algorithms based on significance tests are the SGS algorithm for learning DAG models and its more effective modifications known as the PC algorithm and the PC* algorithm described in [117]. These procedures stem from the premise called the *data faithfulness assumption* which can be paraphrased as follows (cf. Section 1.1):

> data are 'generated' by a probability measure $P$ which is perfectly Markovian with respect to an object $\boldsymbol{o}$ within the considered universum of objects of discrete mathematics.

In the case of the above algorithms from [117] the universum of the objects is the collection of acyclic directed graphs over $N$. The algorithms are based on statistical CI tests, that is, tests which – on the basis of given data – either reject or do not reject the hypothesis that a certain CI statement $a \perp\!\!\!\perp b \,|\, C\ [P]$ is true. Tests of this kind are usually derived from statistics which are known to be the measures of stochastic conditional dependence (see Section 10.10.1 for the concept of a statistic). For example, in § 5.5 of [117] two statistics of this kind are mentioned in the discrete case: the $X^2$-statistics and the $G^2$-statistics. The goal of the algorithms is to determine the equivalence class of graphs consisting of those acyclic directed graphs with respect to which $P$ is perfectly Markovian. The basic step is an observation that if $P$ is perfectly Markovian with respect to an acyclic directed graph $G$, then the following characterization of edges and immoralities in $G$ holds:

$$[a,b] \text{ is an edge in } G \quad \Leftrightarrow \quad \forall\, C \subseteq N \setminus \{a,b\} \quad a \not\perp\!\!\!\perp b \,|\, C\ [P]\,,$$

and if $[a,b]$, $[b,c]$ are edges in $G$ while $[a,b]$ is not an edge in $G$ then

$$a \to c \leftarrow b \text{ in } G \quad \Leftrightarrow \quad \forall\, C\ c \in C \subseteq N \setminus \{a,b\} \quad a \not\perp\!\!\!\perp b \,|\, C\ [P]$$

- see [144] or [56]. On the condition that data are 'generated' from $P$, the above mentioned statistical CI tests give a criterion for the composite conditional dependence statements on the right-hand side. This allows one to determine the underlying graph of $G$ and all immoralities in $G$. Thus, a hybrid graph $H$, called a *pattern*, can be obtained, which has the same underlying graph as $G$ and just those arrows which belong to the immoralities in $G$; the other edges in $H$ are lines. The final step is a purely graphical procedure whose aim is to direct some other edges of the pattern so as to get a direction that is shared by all acyclic directed graphs equivalent to $G$. Nevertheless, the hybrid graph $H_{\ddagger}$ obtained by the algorithm from [117] need not be complete in the sense that every *essential arrow*, that is, an arrow which has the same direction in all acyclic directed graphs equivalent to $G$, is an arrow with this direction in $H_{\ddagger}$. However, the final step of the SGS algorithm may be replaced by the procedure described in [87]. The resulting hybrid graph is then the essential graph, named also the completed pattern, which can serve as the unique representative of the equivalence class of acyclic directed graphs with respect to which $P$ is perfectly Markovian.

Analogous procedures, sometimes called *recovery algorithms*, were proposed for other classes of graphs. An algorithm of this kind for (classic) chain graphs whose result is the largest chain graph (see p. 52) was presented in [133]. In § 5.6 of [55] a reference is given to an algorithm for learning decomposable models which is based on statistical tests and on the idea of gradual decomposition of cliques of a triangulated undirected graph. A modification of the PC algorithm for learning acyclic directed graphs with hidden variables is described in § 6.2 of [117].

All the above-mentioned algorithms are ultimately based on the data faithfulness assumption and the validity of this assumption can hardly be ensured. Indeed, its validity for arbitrary data actually means that, for every probability measure $P$ in the considered distribution framework, there exists an object $\boldsymbol{o}$ of discrete mathematics in the considered universum of objects such that $P$ is perfectly Markovian with respect to $\boldsymbol{o}$. This is nothing but the completeness requirement mentioned in Section 1.1. However, it is explained in Section 3.6 that such a condition cannot be fulfilled for an arbitrary class of graphical models and the discrete distribution framework. The point of this consideration is as follows:

> *learning methods based on statistical CI tests are safely applicable only if the data faithfulness assumption is guaranteed.*

REMARK 8.1 A boon of the book [117] is that the data faithfulness assumption was made explicit there and the limitation of learning methods based on statistical CI tests was indicated in that way. Unfortunately, some researchers in the area of artificial intelligence do not seem to be aware of this limitation, which leads to the methodological error mentioned in the end of Section 1.1. These researchers use the procedures for learning Bayesian networks based on CI tests for 'single edge removals' which have already been criticized in [149]. Examples of similar procedures are the `Construct`-TAN algorithm for 'learning restricted Bayesian networks' presented in [35] and its modifications from [20]. The argument used on p. 8 of [35] to justify the restriction of the class of TAN models is that the model learned by the usual method of the maximization of a quality criterion may result in a 'poor' classification. However, the `Costruct`-TAN algorithm is ultimately based on the values of the $G^2$-statistic for triplets of the form $\langle a, b|c\rangle$, $a, b, c \in N$ which correspond to the CI test for the respective elementary statement $a \perp\!\!\!\perp b\,|\,c$. Owing to the restriction to a special class of TAN models, the 'learned' model always involves a fixed-context CI statement $a \perp\!\!\!\perp b\,|\,N \setminus \{a, b\}$. But no evidence for the validity of this CI statement was provided by data and the procedure is, therefore, methodologically incorrect, unless the data faithfulness assumption is somehow ensured!

Another issue is that researchers in the area of artificial intelligence often support the applicability of their various heuristic algorithms for learning Bayesian networks of the above type by computer simulations. The trouble is that, in simulations of this type, data are 'generated' artificially from a probability measure which is perfectly Markovian with respect to an acyclic directed graph. Thus, the data faithfulness assumption is 'fulfilled' as a result of the way in which the artificial data were generated! These algorithms may appear worthless if one is expected to work with such real data for which this assumption is not satisfied. I think that real data of this sort are quite common. In my view, a very simple example of real data which do not meet the data faithfulness assumption was already given by Whittaker – see Example 8.4.1 and Exercise 1 in § 8.6 in [148]. $\triangle$

### Need for lattice structure

One of the ways to behave in a situation in which data do not fit one of the considered models is the change of the objective of the learning procedure. Instead of seeking 'the only true model' the aim may become to find the 'best approximation' in the universum of considered statistical models or an approximation which is 'good enough'. Then, of course, one has to introduce a suitable measure of approximation.

An example of an algorithm whose aim is to find an approximation in the above sense and which can also be interpreted as an algorithm based on significance tests is the *simple two step procedure* for learning UG models described in § 8.3 of Whittaker's book [148]. The measure Whittaker uses is the *deviance* of a statistical model $\mathbb{M}_G$ determined by an undirected graph $G$ (see Section 10.10.2). It is defined as a multiple by 2 of the difference between the unconstrained maximum and the constrained maximum of the logarithm of the likelihood function over $\mathbb{M}$ (see Section 10.10.4). Whittaker showed that in the cases he deals with, the deviance has information-theoretical interpretation in terms of relative entropy (see [148] Proposition 6.8.1 in the Gaussian case and Proposition 7.4.2 in the discrete case). The procedure described by Whittaker is a series of mutual comparisons of two statistical models $\mathbb{M}$ and $\mathbb{M}'$ which are nested each time. Moreover, during the algorithm only 'close' models are compared, that is, models determined by undirected graphs which differ in the presence of one edge. This means, in the statistical interpretation (see Section 10.10.2), that parameterizations differ in the presence of one parameter. To compare statistical models, a significance test based on the *deviance difference* is used. Whittaker explicitly wrote on p. 228 of [148] that, if possible, he would like to give the meaning to the deviance difference (between two close models), namely the meaning of a statistic for testing (elementary) CI statements. Indeed, this interpretation is possible in most of the situations Whittaker deals with; for example, for a pair of decomposable models in the discrete distribution framework - see Remark 8.14 on p. 161. The 'simple two stage procedure' (p. 256 in [148]) could be described as follows:

**(i)** The algorithm starts with the *saturated model* (see Section 10.10.3), that is, the model determined by an undirected graph $H^0$ in which every pair of distinct nodes in $N$ is a line.

**(ii)** The significance tests for the presence or the absence of each particular edge in $H^0$ are performed. In the discrete case, these are the statistical tests for CI statements of the form $d \perp\!\!\!\perp e \,|\, N \setminus \{d, e\}$ based on the $G^2$-statistics (see Section 10.10.1).

**(iii)** The lines $a - b$ in $H^0$ for which these tests do not reject the hypothesis about the validity of $a \perp\!\!\!\perp b \,|\, N \setminus \{a, b\}$ are removed from $H^0$ and a graph $H^1$ is obtained.

This is the end of the so-called *backward phase* of the procedure. Since the deviance of the model $\mathbb{M}_{H^1}$ may be too high, the algorithm may continue by its *forward phase*:

**(iv)** For every pair $[a, b]$ which represents a missing line in $H^1$, one computes the deviance differences between the model determined by a graph which is created by adding $a - b$ to $H^1$ and the model determined by $H^1$. Note that the result is often the value of the $G^2$-statistics for testing a certain CI statement $a \perp\!\!\!\perp b \,|\, C$ where $C \subseteq N \setminus \{a, b\}$ is a certain set separating $a$ and $b$ in $H^1$.

**(v)** Finally, those lines $a - b$ which correspond to significant deviance differences are added to $H^1$ and the resulting undirected graph $H^2$ is obtained.

Whittaker, who also offered similar procedures for other universa of statistical models, mentioned on p. 245 of [148] that the universa of models he considers often have the *lattice* *structure*. In my view, this is the crucial requirement on the universum of statistical models so that a learning method based on significance tests could be applied. Actually, it is a requirement on the respective class of formal independence models. Indeed, one can describe the backward phase of the above procedure from the point of view of CI interpretation as follows. In **(ii)** certain acceptable UG models were generated and in **(iii)** one moves to their infimum (provided one has in mind the order $\mathcal{M}_G \subseteq \mathcal{M}_H$ between UG models). Moreover, in the forward phase, in **(iv)** some suitable UG models are generated, that is, models which seem better than the 'current' UG model $\mathcal{M}_{H^1}$. Then in **(v)** one moves to their supremum.

Thus, the assumption that the class of induced formal independence models has the lattice structure allows one to combine the knowledge obtained as a result of various particular CI tests. Actually, an analogous situation occurs in the case of learning CI structure on the basis of the knowledge provided by human experts (see Section 1.1). Imagine a situation in which two different experts provide the information about CI structure in the form of two objects of discrete mathematics $\boldsymbol{o}_1$ and $\boldsymbol{o}_2$, for example graphs over $N$. The knowledge can be expressed in the form of the induced formal independence models $\mathcal{M}_{\boldsymbol{o}_1}$ and $\mathcal{M}_{\boldsymbol{o}_2}$ (see Section 2.2.1). If one relies on both experts then one would like to represent the joint knowledge (within the considered class of models) by means of an object $\boldsymbol{o}$ such that $\mathcal{M}_{\boldsymbol{o}}$ involves both $\mathcal{M}_{\boldsymbol{o}_1}$ and $\mathcal{M}_{\boldsymbol{o}_2}$ and $\mathcal{M}_{\boldsymbol{o}}$ is the least possible model of this kind. A natural requirement for the uniqueness of this model is just the requirement of the supremum existence. Well, in the case of a suspicious attitude one would only like to accept the pieces of information which are confirmed by both experts and the natural requirement is the requirement of the infimum existence.

The goal of the above thoughts is the following conclusion:

> *if the data faithfulness assumption is not fulfilled and the class of formal independence models induced by objects from the considered universum of objects of discrete mathematics is not a lattice, then a method based on statistical tests of particular CI statements may appear not to be applicable.*

Unfortunately, some classes of graphical models do not meet the above condition. For example, the class of DAG models for $|N| \geq 3$ is not a lattice; see Figure 7.3 where the model which corresponds to $a \perp\!\!\!\perp b \,|\, \emptyset$ and the model which corresponds to $a \perp\!\!\!\perp b \,|\, c$ have no join but two minimal upper bounds, namely the models which describe $a \perp\!\!\!\perp bc \,|\, \emptyset$ and $b \perp\!\!\!\perp ac \,|\, \emptyset$. Another example of a class of models which is not a lattice is the class of decomposable models for $|N| \geq 4$ (c.f., the example on p. 254 of [148]).

Remark 8.2 Whittaker gave examples showing that his simple two step procedure both may and may not lead to the choice of an adequate statistical model, that is, a model with sufficiently low deviance (Examples 8.5.1 and 8.4.2 in [148]). In my view, failing in the aim of the procedure is caused by the fact that the data faithfulness assumption is not 'valid' for UG models in that case. The above-mentioned procedures and other procedures mentioned in § 8.3 of [148] can alternatively be understood as modifications of methods based on the maximization of a quality criterion from Section 8.2. Indeed, the deviance of a statistical model $\mathbb{M}_H$ for an undirected graph $H$ has the form $\boldsymbol{k} - 2 \cdot \mathsf{MLL}\,(\mathbb{M}_H, D)$ where $\boldsymbol{k}$ is a constant and $\mathsf{MLL}\,(\mathbb{M}_H, D)$ is the value of maximized log-likelihood criterion for the model $\mathbb{M}_H$ and data $D$ (see Section 10.10.4). Thus, the task of finding a model which has the deviance as low as possible is equivalent to the task of finding a model which has the value of $\mathsf{MLL}\,(\mathbb{M}_H\,, D)$ as high as possible.

In Example 8.2.2 [148] Whittaker mentioned an interesting phenomenon: the deviance difference for different pairs of UG models may coincide! Moreover, the deviance difference is then the value of the $G^2$-statistic for testing a certain CI statement. This feature could be nicely explained as a special case of Observation 8.4 - see Remark 8.14. $\triangle$

**Universum of structural imsets**

The goal of the considerations in this section, which are to a certain extent an extension of the motivation account from Section 1.1, is the moral that in the case of learning methods based on statistical CI tests, the universum of considered objects of discrete mathematics should satisfy two natural conditions. The first condition is the *completeness* with respect to the respective distribution framework and the second condition is the requirement that the class of induced formal independence models has the lattice structure.

The universum of structural imsets satisfies both of these conditions. The completeness with respect to the discrete and Gaussian distribution framework follows from Theorem 5.2; the fact that the class of structural models is a lattice follows from Theorem 5.3. If one dreams about a learning procedure for structural imsets based on statistical CI tests, then baricentral imsets considered in Section 7.1 are perhaps the most suitable representatives of classes of (facially) equivalent structural imsets.

## 8.2 Quality criteria

Most of the algorithms for learning graphical models are based on the idea of maximizing a *quality criterion* which is a function ascribing to a graph and data a real number which 'evaluates' how the statistical model described by the graph fits the data. Alternative names for a quality criterion are 'quality measure' [13], 'score criterion' [24] and 'score metric' [22]. Maximization of a quality criterion is often used in the area of artificial intelligence and seems to be more suitable for learning DAG models and decomposable models than methods based on significance tests. One can distinguish at least three methodological approaches to derivation of quality criteria (for details see Section 4.2 in [13]).

- Classic statistical interpretation of a graphical model in terms of a parameterized class of probability measures leads to *information criteria*; for example, the AIC criterion and the BIC criterion - see Section 10.10.4.

- Bayesian approach is based on an additional assumption that a 'prior' probability measure on the respective parameter space is given for each considered graphical model. This approach, initiated by the paper [23], leads to a variety of *Bayesian criteria* (see § 3.2 of [19]).

- *Minimum description length* principle [61] stems from coding theory. Roughly said, the basic idea is to evaluate a graphical model by the overall number of bits needed to encode data within the model. However, the resulting MDL criterion appears to coincide with the classic BIC information criterion (see § 4.2.3 of [13]).

The weak point of this approach is that the class of considered graphs is often too large, which makes the direct maximization of a quality criterion infeasible. To avoid this problem, various heuristic *local search methods* were developed (see § 4.3 in [13]). Every method of this kind has its specific *search space*, which consists of the collection of *states* and the collection of *moves* between the states. The states are typically graphs over $N$, either acyclic directed graphs or essential graphs, and the moves are some predefined 'local' graphical operations with them. Thus, in order to apply a local search method one needs to introduce a *neighbourhood structure* in the class of considered graphs. Every graph is assigned a relatively small set of neighbouring graphs, which are typically graphs differing from the considered graph in the presence of one edge. The point is that most quality criteria used in practice have a pleasant property that the difference of their values for neighbouring graphs is a single term which is easy to compute - see Section 8.2.3. Thus, instead of the 'global' maximum of a quality criterion, various algorithms allow us to find a 'local' maximum relative to the neighbourhood structure in consideration. On the other hand, some algorithms of this kind are guaranteed to achieve the global maximum provided that certain quite strong conditions are fulfilled (these conditions include the data faithfulness assumption) - see [89].

Remark 8.3 Some of the Bayesian approaches come even with a further assumption that a 'prior' probability measure on the finite class of considered graphs is given. This assumption is then reflected by an additional term in the respective Bayesian criterion. Note that this additional assumption does not seem to prevent one from using local search methods.

Nevertheless, the whole collection of assumptions – that is, every graphical model is understood as a parameterized class of probability measures, a certain reasonable system of 'priors' on parameter spaces is given and a 'prior' on the class of models is specified – allows one to apply a special Bayesian simulation method known as the *Markov chain Monte Carlo* (MCMC) method. From a certain perspective, this method can be viewed as a particular method for learning graphical models which is based neither on significance tests nor on the maximization of a quality criterion. The idea is that the above assumptions allow one to define, on the basis of data, a 'posterior' probability measure on the class of considered graphs and a system of 'posteriors' on parameter spaces. The posteriors could be interpreted as stochastic estimates of both the graphical models and its parameters. However, their direct computation is again infeasible. To overcome this obstacle, a stochastic simulation method is used, whose aim is to get an approximation of these 'posteriors'. Of course, lots of technical assumptions are accepted (which are intentionally omitted here) to make the method feasible. Nevertheless, various versions of the MCMC method are similar to the local search methods mentioned above; they are also based on the idea of search space and the idea of 'travelling' in the class of considered graphs - for details see § 3.6 in [19]. Some theoretical results ensure the (stochastic) convergence of the MCMC method if mild assumptions are fulfilled. The method was used both for learning decomposable models [40] and DAG models [73]. △

### 8.2.1 Criteria for learning DAG models

The aim of this section is to provide details about quality criteria for learning DAG models in the case of the discrete distribution framework (see Section 10.10.3). First, let us introduce the symbol $\mathsf{DAGS}(N)$ to denote the collection of all acyclic directed graphs having $N$ as the set of nodes. Let a discrete joint sample space $\mathsf{X}_N = \prod_{i\in N} \mathsf{X}_i$ over $N$ (see Section 10.10.1) be fixed. By a *quality criterion* for learning DAG models adapted to this sample space any function $\mathcal{Q}$ will be understood which ascribes a real number $\mathcal{Q}(G, D)$ to a graph $G \in \mathsf{DAGS}(N)$ and a database $D \in \mathsf{DATA}(N, d)$ (see p. 202 for this notation).

To derive formulas for basic information criteria details about the parameterization of statistical models described by acyclic directed graphs are needed. The statistical model described by $G \in \mathsf{DAGS}(N)$ consists of the class $\mathbb{M}_G$ of probability measures on $\mathsf{X}_N$ which are Markovian with respect to $G$ (see p. 47). Recall that a probability measure $P$ on $\mathsf{X}_N$ is uniquely determined by its density $f$ with respect to the counting measure $\upsilon$ on $\mathsf{X}_N$, that is, $f(x) = P(\{x\})$ for every $x \in \mathsf{X}_N$. The respective *marginal densities* of $P$ (see Convention 1 on p. 25) are then

$$f_A(y) = P^A(\{y\}) = \sum_{z\in \mathsf{X}_{N\setminus A}} f(y, z) \quad \text{for } \emptyset \neq A \subset N,\ y \in \mathsf{X}_A\,,$$

and $f_N \equiv f$, $f_\emptyset \equiv 1$ by a convention. The *conditional density* $f_{A|C}$ of $P$ for disjoint $A, C \subseteq N$ can be defined as follows:

$$f_{A|C}(x|z) = \begin{cases} \frac{f_{AC}([x,z])}{f_C(z)} & \text{if } f_C(z) = P^C(\{z\}) > 0, \\ 0 & \text{if } f_C(z) = 0 \text{ by convention,} \end{cases} \qquad \text{for } x \in \mathsf{X}_A,\ z \in \mathsf{X}_C\,.$$

The well-known fact (see Theorem 1 in [67]) is that $P \in \mathbb{M}_G$ iff it is *recursively factorizable* according to $G$, that is,

$$f(x) = \prod_{i\in N} f_{i|pa_G(i)}(x_i|x_{pa_G(i)}) \qquad \text{for every } x \in \mathsf{X}_N\,. \tag{8.1}$$

Note that this fact can also be derived as a consequence of Theorem 4.1: owing to Lemma 7.1, it suffices to apply the theorem to the standard imset $u_G$ and the standard reference system (in the discrete case) - see p. 69. The definition of conditional density then allows us to rewrite the condition **(i)** in Theorem 4.1 as the formula (8.1).

Convention 3 Let $\mathsf{X}_N = \prod_{i\in N} \mathsf{X}_i$ be a discrete sample space over $N$. The letter $i$ will be used as *a generic symbol for a variable in* $N$: $i \in N$.

Denote $r(i) = |\mathsf{X}_i| \geq 1$ and fix an ordering $y_i^1, \ldots, y_i^{r(i)}$ of elements of $\mathsf{X}_i$ for every $i \in N$. The letter $k$ will be used as *a generic symbol for a code of an element of* $\mathsf{X}_i$: $k \in \{1, \ldots, r(i)\}$. More specifically, $k$ is the code of the $k$-th configuration $y_i^k$ in the fixed ordering. Given $i \in N$ and $x \in \mathsf{X}_A$ such that $i \in A \subseteq N$ symbol $k(i, x)$ will denote the code of $x_i$, that is, the unique $1 \leq k \leq r(i)$ such that $x_i = y_i^k$.

Let $G \in \mathsf{DAGS}(N)$. Denote by $q(i, G) \equiv |\mathsf{X}_{pa_G(i)}| = \prod_{l\in pa_G(i)} r(l) \geq 1$ the number of *parent configurations* for the variable $i \in N$ and accept the convention that $q(i, G) = 1$ if $pa_G(i) = \emptyset$. Fix an ordering $z_i^1, \ldots, z_i^{q(i,G)}$ of parent configurations, that is, elements of $\mathsf{X}_{pa_G(i)}$, for every $i \in N$. If $pa_G(i) = \emptyset$ then it consists of the only parent configuration. The letter $j$ will be used as *a generic symbol for a code of a parent configuration*: $j \in \{1, \ldots, q(i, G)\}$. Thus, $j$ is the code of the $j$-th configuration $z_i^j$ in the fixed ordering. Given $i \in N$ and $x \in \mathsf{X}_A$ such that $pa_G(i) \subseteq A \subseteq N$ the symbol $j(i, x)$ will denote the code of $x_{pa_G(i)}$, that is, the unique $1 \leq j \leq q(i,G)$ such that $x_{pa_G(i)} = z_i^j$ if $pa_G(i) \neq \emptyset$ and $j = 1$ otherwise.

Moreover, let $D \in \mathsf{DATA}(N,d)$, $d \in \mathbb{N}$ be data over $N$ (see p. 202); more specifically $D : x^1, \ldots, x^d$. Introduce the following special notation for the numbers of configuration occurrences in the database $D$:

$$\begin{aligned} d_{ij} &= |\{1 \leq \ell \leq d;\ x^\ell_{pa_G(i)} = z_i^j\}| \quad \text{for } i \in N,\, j \in \{1,\ldots,q(i,G)\}, \\ d_{ijk} &= |\{1 \leq \ell \leq d;\ x^\ell_{\{i\}\cup pa_G(i)} = (y_i^k, z_i^j)\}| \quad \text{for } i \in N,\, j \in \{1,\ldots,q(i,G)\},\, k \in \{1,\ldots,r(i)\}, \\ d_{\boldsymbol{y}} &= |\{1 \leq \ell \leq d;\ x^\ell_A = \boldsymbol{y}\}| \quad \text{for } \emptyset \neq A \subseteq N,\, \boldsymbol{y} \in \mathsf{X}_A\,. \end{aligned}$$

Of course, $d_{i1} = d$ if $pa_G(i) = \emptyset$.

Given $G \in \mathsf{DAGS}(N)$, the respective 'standard' parameterization of $\mathbb{M}_G$ is based on the recursive factorization (8.1). The set of parameters $\Theta_G$ consists of vectors

$$\begin{gathered} \boldsymbol{\theta} \,:\, \theta_{ijk} \in [0,1] \quad \text{for } i \in N,\, j \in \{1,\ldots,q(i,G)\},\, k \in \{1,\ldots,r(i)\} \\ \text{such that } \sum_{k=1}^{r(i)} \theta_{ijk} = 1 \text{ for every } i \in N,\ 1 \leq j \leq q(i,G)\,. \end{gathered} \tag{8.2}$$

Actually, every parameter $\theta_{ijk}$ can be interpreted as the value of the conditional density $f_{i|pa(i)}(y_i^k|z_i^j)$. Given a vector parameter $\boldsymbol{\theta}$, the respective probability measure on $\mathsf{X}_N$ is given by its density

$$f^{\boldsymbol{\theta}}(x) \;=\; \prod_{i\in N} \theta_{i\,j(i,x)\,k(i,x)} \qquad \text{for } x \in \mathsf{X}_N\,. \tag{8.3}$$

Lemma 8.1 Let us accept Convention 3 and suppose that a set of parameters $\boldsymbol{\theta}$ satisfying (8.2) is given. Then the formula (8.3) defines a density of a probability measure $P^{\boldsymbol{\theta}}$ on $\mathsf{X}_N$ (with respect to the counting measure). Moreover, $P^{\boldsymbol{\theta}} \in \mathbb{M}_G$ and $\forall\, i \in N\ \forall j \in \{1,\ldots,q(i,G)\}\ \forall\, k \in \{1,\ldots,r(i)\}$

$$f^{\boldsymbol{\theta}}_{i|pa_G(i)}(y_i^k|z_i^j) \;=\; \theta_{ijk} \qquad \text{if } f^{\boldsymbol{\theta}}_{pa_G(i)}(z_i^j) > 0\,. \tag{8.4}$$

The mapping $\boldsymbol{\theta} \mapsto P^{\boldsymbol{\theta}}$ is a mapping onto $\mathbb{M}_G$.

**Proof:** Given $A \subseteq N$ let us denote the collection of parameters $\theta_{ijk}$ with $i \in A$ by $\boldsymbol{\theta}[A]$. Let $t \in N$ be a terminal node in $G$. The first observation is that

$$f^{\boldsymbol{\theta}}([y,z]) \;=\; (\prod_{i\in N\setminus t} \theta_{i\,j(i,z)\,k(i,z)}) \cdot \theta_{t\,j(t,z)\,k(t,y)} \quad \text{for every } z \in \mathsf{X}_{N\setminus t},\ y \in \mathsf{X}_t\,. \tag{8.5}$$

Indeed, given $x = [y,z] \in \mathsf{X}_N$ by Convention 3 and the fact that $pa_G(i) \subseteq N \setminus \{t\}$ for $i \in N$ one has $j(i,x) = j(i,z)$ for $i \in N$, $k(t,x) = k(t,y)$ and $k(i,x) = k(i,z)$ for $i \in N \setminus \{t\}$. It suffices to substitute these equalities into (8.3).
The second observation is that

$$f^{\boldsymbol{\theta}}_{N\setminus t}(z) \equiv \sum_{y\in \mathsf{X}_t} f^{\boldsymbol{\theta}}([y,z]) = \prod_{i\in N\setminus t} \theta_{i\,j(i,z)\,k(i,z)} = f^{\boldsymbol{\theta}[N\setminus t]}(z) \quad \text{for every } z \in \mathsf{X}_{N\setminus t}\,. \tag{8.6}$$

Indeed, substitute (8.5) in $f^{\boldsymbol{\theta}}_{N\setminus t}(z) = \sum_{y\in\mathsf{X}_t} f^{\boldsymbol{\theta}}([y,z])$ and get $f^{\boldsymbol{\theta}}_{N\setminus t}(z) = (\prod_{i\in N\setminus t} \theta_{i\,j(i,z)\,k(i,z)}) \cdot \sum_{y\in\mathsf{X}_t} \theta_{t\,j(t,z)\,k(t,y)}$. The last sum is 1 by (8.2) and one can apply (8.3) to $G_{N\setminus t}$ and $\boldsymbol{\theta}[N \setminus t]$.

To verify that $f^{\boldsymbol{\theta}}$ is a probability density, choose an ordering $t_1, \ldots, t_{|N|}$ of the nodes of $G$ which is consonant with the direction of arrows, that is, $t_n \to t_m$ in $G$ implies $n < m$. Put $N(n) = \{t_1, \ldots, t_n\}$ for $n = 1, \ldots, |N|$, observe that $t_n$ is a terminal node in $G_{N(n)}$ and prove by induction on $n = 1, \ldots, |N|$ that $f^{\boldsymbol{\theta}[N(n)]}$ is a probability density on $\mathsf{X}_{N(n)}$. Indeed, it is evident that $f^{\boldsymbol{\theta}[N(n)]}(x) \geq 0$ for $x \in \mathsf{X}_{N(n)}$. The fact $\sum_{x \in \mathsf{X}_{N(n)}} f^{\boldsymbol{\theta}[N(n)]}(x) = 1$ for $n = 1$ is implied by (8.2) and the induction step follows from (8.6).
Thus, $f^{\boldsymbol{\theta}}$ defines a probability measure $P^{\boldsymbol{\theta}}$ on $\mathsf{X}_N$. The next observation is that, provided $t$ is a terminal node in $G$, one has

$$f^{\boldsymbol{\theta}}_{t\,|\,N\setminus t}(y|z) \;=\; \theta_{t\,j(t,z)\,k(t,y)} \quad \text{for } z \in \mathsf{X}_{N\setminus t} \text{ with } f^{\boldsymbol{\theta}}_{N\setminus t}(z) > 0,\; y \in \mathsf{X}_t\,. \tag{8.7}$$

Indeed, substitute (8.5) and (8.6) into the definition of $f^{\boldsymbol{\theta}}_{t\,|\,N\setminus t}(y|z)$. The term $\prod_{i \in N\setminus t} \theta_{i\,j(i,z)\,k(i,z)}$ is ensured to be non-zero in this case. Thus, it can be cancelled.
It follows from (8.5) by means of Lemma 2.5 that $t \perp\!\!\!\perp N \setminus t \cup pa_G(t) \,|\, pa_G(t)\ [P^{\boldsymbol{\theta}}]$. An analogous argument applied to $t_n$ and $G_n$ gives $t_n \perp\!\!\!\perp N(n-1) \setminus pa_G(t_n) \,|\, pa_G(t_n)\ [P^{\boldsymbol{\theta}}]$ for $n = 1, \ldots, |N|$. This property, known as the 'local Markov property' (see Remark 3.4) implies by the result from [143] that $P^{\boldsymbol{\theta}}$ is (globally) Markovian with respect to $G$, that is, $P^{\boldsymbol{\theta}} \in \mathbb{M}_G$ (see also Theorem 1 in [67]). The next observation is that if $t$ is a terminal node in $G$ then

$$f^{\boldsymbol{\theta}}_{t\,|\,pa_G(t)}(y^k_t|z^j_t) \;=\; \theta_{tjk} \quad \text{for } j \in \{1, \ldots, q(t,G)\} \text{ with } f^{\boldsymbol{\theta}}_{pa_G(t)}(z^j_t) > 0,\; k \in \{1, \ldots, r(t)\}\,.$$

Indeed, choose $z \in \mathsf{X}_{N\setminus t}$ such that $f^{\boldsymbol{\theta}}_{N\setminus t}(z) > 0$ and $z_{pa(t)} = z^j_t$, that is, $j = j(t,z)$. Then take $y \in \mathsf{X}_t$ such that $k = k(t,y)$. Thus, one is entitled to apply (8.7) to $z$. Write $f^{\boldsymbol{\theta}}_{t\,|\,pa(t)}(y^k_t|z^j_t) = f^{\boldsymbol{\theta}}_{t\,|\,pa(t)}(y|z_{pa(t)}) = (f^{\boldsymbol{\theta}}_{pa(t)}(z_{pa(t)}))^{-1} \cdot f^{\boldsymbol{\theta}}_{t \cup pa(t)}([y, z_{pa(t)}])$. However, $t \perp\!\!\!\perp N \setminus t \cup pa(t) \,|\, pa(t)\ [P^{\boldsymbol{\theta}}]$ implies by Lemma 2.4 that the latter term equals $(f^{\boldsymbol{\theta}}(z))^{-1} \cdot f^{\boldsymbol{\theta}}([y,z])$. By (8.7), it equals $\theta_{t\,j(t,z)\,k(t,y)} = \theta_{tjk}$.
The same argument applied to $t_n$ and $G_{N(n)}$ for $n = 1, \ldots, |N|$ gives (8.4). Finally, given $P \in \mathbb{M}_G$ let us put $\theta_{ijk} = f_{i\,|\,pa_G(i)}(y^k_i|z^j_i)$ for every $i \in N$, $j \in \{1, \ldots, q(i,G)\}$, $k \in \{1, \ldots, r(i)\}$. Observe that (8.2) is valid, which allows one to define $P^{\boldsymbol{\theta}}$. By (8.1) and (8.3) get $f(x) = \prod_{i \in N} \theta_{i\,j(i,x)\,k(i,x)} = f^{\boldsymbol{\theta}}(x)$ for every $x \in \mathsf{X}_N$, which implies $P = P^{\boldsymbol{\theta}}$. $\square$

Remark 8.4 The correspondence $\boldsymbol{\theta} \mapsto P^{\boldsymbol{\theta}}$ is not a one-to-one mapping. The reason is that if $t$ is a terminal node in $G$ and $P^{\boldsymbol{\theta}}_{N\setminus t}(x_{N\setminus t}) = 0$ then $P^{\boldsymbol{\theta}}(x) = 0$ no matter what is the value of $\theta_{t\,j(t,x)\,k(t,x)}$. However, it is a one-to-one mapping on the set of parameters $\boldsymbol{\theta}$ such that $\theta_{ijk} \in (0,1)$ for every $i, j, k$. These parameters correspond to strictly positive probability measures on $\mathbb{M}_G$. $\triangle$

To derive a maximum likelihood estimate in $\mathbb{M}_G$ and the formula for the $\mathsf{MLL}$ quality criterion the following technical lemma is needed.

Lemma 8.2 Let $d_k \geq 0$, $k \in K$ be a non-empty finite collection of non-negative numbers such that $d = \sum_{k \in K} d_k > 0$. Then the vector function $f$ with values in $[-\infty, \infty)$ given by

$$f([\theta_k]_{k \in K}) \;=\; \sum_{k \in K} d_k \cdot \ln \theta_k\,,$$

defined on the domain $\mathcal{D}(f) = \{[\theta_k]_{k \in K}\,;\ \theta_k \geq 0\ \ \sum_{k \in K} \theta_k = 1\}$ attains its maximum in $[\hat{\theta}_k]_{k \in K}$ where $\hat{\theta}_k = \frac{d_k}{d}$ for $k \in K$. In particular, the maximum value of $f$ on $\mathcal{D}(f)$ is $\sum_{k \in K, d_k > 0} d_k \cdot \ln \frac{d_k}{d}$.

**Proof:** The first step is the following auxiliary observation.
**I.** Suppose $0 \leq s \leq 1$, $d_i, d_j \geq 0$ and $d_i + d_j > 0$. Then the function

$$g(\theta_i, \theta_j) \;=\; d_i \cdot \ln \theta_i + d_j \cdot \ln \theta_j \qquad \text{defined for } \theta_i, \theta_i \geq 0,\ \theta_i + \theta_j = 1 - s\,,$$

taking values in $[-\infty, \infty)$ attains its maximum at $(\hat{\theta}_i, \hat{\theta}_j) = (\frac{d_i}{d_i+d_j} \cdot (1-s), \frac{d_j}{d_i+d_j} \cdot (1-s))$. Indeed, the claim is trivial for $s = 1$. If $s < 1$ then $\theta_j = 1 - s - \theta_i$ and the task is equivalent to the maximization of the function $\tilde{g}(\theta) = d_i \cdot \ln \theta + d_j \cdot \ln(1 - s - \theta)$ for $\theta \in [0, 1-s]$. The rest is elementary computation.
Lemma 8.2 can be verified by induction on $|K|$. It is trivial if $|K| = 1$; if $|K| = 2$ then it is implied by Step **I** with $K = \{i, j\}$ and $s = 0$. To verify the induction step suppose $|K| \geq 3$, choose distinct $i, j \in K$ such that $d_i + d_j > 0$ and put $L = K \setminus \{i, j\}$. Introduce notation $\theta_L = [\theta_l]_{l \in L}$, $s(\theta_L) = \sum_{l \in L} \theta_l$ and put

$$\varphi_i(\theta_L) = \frac{d_i}{d_i + d_j} \cdot (1 - s(\theta_L)), \quad \varphi_j(\theta_L) = \frac{d_j}{d_i + d_j} \cdot (1 - s(\theta_L))\,,$$

for every $\theta_L$ with $\theta_l \geq 0$, $l \in L$ and $s(\theta_L) \leq 1$. Observe that $[\theta_L, \varphi_i(\theta_L), \varphi_j(\theta_L)] \in \mathcal{D}(f)$. Moreover, for $\hat{\theta}_L = [\hat{\theta}_l]_{l \in L} = [\frac{d_l}{d}]_{l \in L}$ one has $\varphi_i(\hat{\theta}_L) = \hat{\theta}_i$ and $\varphi_j(\hat{\theta}_L) = \hat{\theta}_j$. Further steps are as follows.
**II.** Observe $f(\theta_K) \leq f(\theta_L, \varphi_i(\theta_L), \varphi_j(\theta_L))$ for every $\theta_K = [\theta_k]_{k \in K} \in \mathcal{D}(f)$.
Indeed, apply Step **I** to $d_i, d_j$ and $s = s(\theta_L)$ which gives $g(\theta_i, \theta_j) \leq g(\varphi_i(\theta_L), \varphi_j(\theta_L))$. Then add $\sum_{l \in L} d_l \cdot \ln \theta_l \in [-\infty, 0]$ to both sides of the obtained inequality to get what is desired.
**III.** Observe that $f(\theta_L, \varphi_i(\theta_L), \varphi_j(\theta_L)) \leq f(\hat{\theta}_L, \varphi_i(\hat{\theta}_L), \varphi_j(\hat{\theta}_L))$ for every $\theta_L$ with $\theta_l \geq 0$ and $s(\theta_L) \leq 1$.
Let us introduce a new artificial index $r \notin K$ and put $\bar{K} = L \cup \{r\}$, $d_r = d_i + d_j$. Observe that $d = \sum_{k \in \bar{K}} d_k$ and introduce a function $\bar{f}$ taking values in $[-\infty, \infty)$ by

$$\bar{f}([\theta_k]_{k \in \bar{K}}) \;=\; \sum_{k \in \bar{K}} d_k \cdot \ln \theta_k \quad \text{on } \mathcal{D}(f) = \{\, [\theta_k]_{k \in \bar{K}}\,;\, \theta_k \geq 0,\ \sum_{k \in \bar{K}} \theta_k = 1 \,\}\,.$$

By the induction hypothesis, $\bar{f}$ attains its maximum at $\hat{\theta}_{\bar{K}} = [\hat{\theta}_k]_{k \in \bar{K}} = [\frac{d_k}{d}]_{k \in \bar{K}}$. Because of $(\theta_L, 1 - s(\theta_L)) \in \mathcal{D}(\bar{f})$ one has $\bar{f}(\theta_L, 1 - s(\theta_L)) \leq \bar{f}(\hat{\theta}_{\bar{K}})$. But, owing to $d = \sum_{k \in \bar{k}} d_k$, $\hat{\theta}_{\bar{K}} = (\hat{\theta}_L, 1 - s(\hat{\theta}_L))$, which gives $\bar{f}(\theta_L, 1 - s(\theta_L)) \leq \bar{f}(\hat{\theta}_L, 1 - s(\hat{\theta}_L))$. By simple computation observe that

$$f(\theta_L, \varphi_i(\theta_L), \varphi_j(\theta_L)) = \bar{f}(\theta_L, 1 - s(\theta_L)) + d_i \cdot \ln \theta_i + d_j \cdot \ln \theta_j - d_r \cdot \ln \theta_r$$

for any $\theta_L$ in the respective domain, in particular for $\hat{\theta}_L$. Since the constant $d_i \cdot \ln \theta_i + d_j \cdot \ln \theta_j - d_r \cdot \ln \theta_r$ is a finite number, one can add it to the last inequality to get $f(\theta_L, \varphi_i(\theta_L), \varphi_j(\theta_L)) \leq f(\hat{\theta}_L, \varphi_i(\hat{\theta}_L), \varphi_j(\hat{\theta}_L))$.
**IV.** Because $\varphi_i(\hat{\theta}_L) = \hat{\theta}_i$ and $\varphi_j(\hat{\theta}_L) = \hat{\theta}_j$, we have $(\hat{\theta}_L, \varphi_i(\hat{\theta}_L), \varphi_j(\hat{\theta}_L)) = \hat{\theta}_K \in \mathcal{D}(f)$ and using **II** and **III** we get

$$f(\theta_K) \leq f(\theta_L, \varphi_i(\theta_L), \varphi_j(\theta_L)) \leq f(\hat{\theta}_L, \varphi_i(\hat{\theta}_L), \varphi_j(\hat{\theta}_L)) = f(\hat{\theta}_K)$$

for every $\theta_L \in \mathcal{D}(f)$, which concludes the proof of the induction step. □

Consequence 8.1 Let us accept Convention 3. Then the formulas for parameters

$$\hat{\theta}_{ijk} \;=\; \begin{cases} \frac{d_{ijk}}{d_{ij}} & \text{if } d_{ij} > 0, \\ 0 & \text{if } d_{ij} = 0, \end{cases} \qquad \text{for } i \in N,\ 1 \leq j \leq q(i, G),\ 1 \leq k \leq r(i), \tag{8.8}$$

define a maximum likelihood estimate $\hat{P} = P^{\hat{\boldsymbol{\theta}}}$ in $\mathbb{M}_G$ on the basis of data $D \in \mathsf{DATA}(N,d)$ (see Section 10.10.4). Moreover, the maximized log-likelihood criterion has the form

$$\mathsf{MLL}\,(\mathbb{M}_G, D) \;=\; \sum_{i\in N} \sum_{j=1}^{q(i,G)} \sum_{k=1}^{r(i)} d_{ijk} \cdot \ln \frac{d_{ijk}}{d_{ij}} \quad \text{for } G \in \mathsf{DAGS}(N),\ D \in \mathsf{DATA}(N,d), \tag{8.9}$$

where the convention $0 \cdot \ln \frac{0}{*} \equiv 0$ is accepted.

**Proof:** Let $D : x^1, \ldots x^d$, $d \geq 1$. Consider the task of maximizing the logarithm of the likelihood function $l(\theta) = \ln L(\theta, D)$. By the definition of the likelihood function and (8.3) write

$$l(\theta) = \ln L(\theta, D) = \sum_{\ell=1}^{d} \ln f^{\boldsymbol{\theta}}(x^\ell) = \sum_{\ell=1}^{d} \sum_{i\in N} \ln \theta_{i\, j(i,x^\ell)\, k(i,x^\ell)}\,.$$

To get a better formula, introduce auxiliary notation for $i \in N$, $x \in \mathsf{X}_N$, $1 \leq j \leq q(i,G)$ and $1 \leq k \leq r(i)$:

$$\delta_i(j,k,x) = \begin{cases} 1 & \text{if } x_{i\cup pa_G(i)} = (y_i^k, z_i^j), \\ 0 & \text{otherwise.} \end{cases}$$

This allows us to write

$$l(\theta) \;=\; \sum_{\ell=1}^{d} \sum_{i\in N} \sum_{j=1}^{q(i,G)} \sum_{k=1}^{r(i)} \ln \theta_{ijk} \cdot \delta_i(j,k,x^\ell) \;=\; \sum_{i\in N} \sum_{j=1}^{q(i,G)} \sum_{k=1}^{r(i)} \ln \theta_{ijk} \cdot \sum_{\ell=1}^{d} \delta_i(j,k,x^\ell)\,.$$

By Convention 3 we observe that the latter sum is nothing but $d_{ijk}$. Therefore

$$l(\theta) = \sum_{i\in N} \sum_{j=1}^{q(i,G)} \sum_{k=1}^{r(i)} d_{ijk} \cdot \ln \theta_{ijk}\,. \tag{8.10}$$

Clearly, the task of maximizing $l(\theta)$ is equivalent to the task of maximizing, for every $i \in N$ and $j \in \{1, \ldots, q(i,G)\}$, the function $\sum_{k=1}^{r(i)} d_{ijk} \cdot \ln \theta_{ijk}$ defined on the set $\{[\theta_{ijk}]_{k=1}^{r(i)}\,;\ \theta_{ijk} \geq 0\ \sum_{k=1}^{r(i)} \theta_{ijk} = 1\,\}$. Since $\sum_{k=1}^{r(i)} d_{ijk} = d_{ij}$ by Lemma 8.2, if $d_{ij} > 0$ then the maximum is attained at $\hat{\theta}_{ijk} = \frac{d_{ijk}}{d_{ij}}$. If $d_{ij} = 0$ then the function is constant and its maximum 0 is attained at any point, in particular, at $\hat{\theta}_{ijk} = 0$. This gives (8.8). Finally, substitute this into (8.10) to get (8.9). □

Consequence 8.2 Let us accept Convention 3. Then the effective dimension of the statistical model $\mathbb{M}_G$ (see Section 10.10.4) is

$$\mathsf{DIM}\,(G) = \mathsf{DIM}\,(\mathbb{M}_G) = \sum_{i\in N} (r(i)-1) \cdot q(i,G) \;=\; \sum_{i\in N} (r(i)-1) \cdot \prod_{l\in pa_G(i)} r(l)\,. \tag{8.11}$$

In particular, Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) are given by the formulas

$$\mathsf{AIC}\,(G,D) \;=\; \sum_{i\in N} \sum_{j=1}^{q(i,G)} \sum_{k=1}^{r(i)} d_{ijk} \cdot \ln \frac{d_{ijk}}{d_{ij}} - (r(i)-1) \cdot q(i,G) \tag{8.12}$$

$$\mathsf{BIC}\,(G,D) \;=\; \sum_{i\in N} \sum_{j=1}^{q(i,G)} \sum_{k=1}^{r(i)} d_{ijk} \cdot \ln \frac{d_{ijk}}{d_{ij}} - \frac{\ln d}{2} \cdot (r(i)-1) \cdot q(i,G) \tag{8.13}$$

for every $G \in \mathsf{DAGS}(N)$ and every $D \in \mathsf{DATA}(N,d)$.

**Proof:** It follows from (8.2) that, for fixed $i \in N$, $1 \leq j \leq q(i,G)$, the number of linearly independent parameters is $r(i) - 1$. This implies (8.11). The other formulas follow from Consequence 8.1 - see Section 10.10.4. $\square$

Remark 8.5 Bayesian criteria for learning DAG models in a discrete distribution framework could be introduced in the following manner. In general, a prior probability measure $\pi_G$ on the respective parameter space $\Theta_G$ defined by (8.2) is given for every $G \in \mathsf{DAGS}(N)$. This allows one to define the *marginal likelihood* as the integral of the likelihood function with respect to $\pi_G$. The respective Bayesian criterion is the *logarithm of the marginal likelihood* given by

$$\mathsf{LML}\,(G,D) \;=\; \ln \int_{\Theta_G} L(\theta, D)\, d\pi_G(\theta)\,. \tag{8.14}$$

However, to get a more suitable formula for this criterion, additional assumptions are accepted, namely, that every $\pi_G$ is a product measure $\prod_{i\in N}\prod_{j=1}^{q(i,G)} \pi_{(ij)}$ where $\pi_{(ij)}$ is a 'prior' probability measure on the 'local' parameter space $\Theta_{(ij)} = \{[\theta_{ijk}]_{k=1}^{r(i)}\,;\, \theta_{ijk} \geq 0\ \sum_{k=1}^{r(i)} \theta_{ijk}\,\}$ – see the assumptions of 'global' and 'local' parameter independence mentioned in [116]. These assumptions allow one to derive a more explicit formula using (8.10) and the Fubini theorem:

$$\mathsf{LML}\,(G,D) \;=\; \sum_{i\in N}\sum_{j=1}^{q(i,G)} \ln \int_{\Theta_{(ij)}} \prod_{k=1}^{r(i)} \theta_{ijk}^{d_{ijk}}\, d\pi_G(\theta_{ijk})\,. \tag{8.15}$$

Typically, there are lots of technical assumptions on the priors $\pi_{(ij)}$ which allow one to obtain even more convenient formulas for the LML criterion. The flexibility in the choice of priors leads to a great variety of particular Bayesian criteria for learning DAG models. However, because it is not the aim of this monograph to explain peculiarities of the Bayesian approach to learning DAG models, and the reader is referred to § 3.2 of [19]. $\triangle$

### 8.2.2 Score equivalent criteria

A quality criterion $\mathcal{Q}$ for learning DAG models is *score equivalent* if for every pair $G, H \in \mathsf{DAGS}(N)$ and every $D \in \mathsf{DATA}(N,d)$

$$\mathcal{Q}(G,D) \;=\; \mathcal{Q}(H,D) \qquad \text{whenever } G \text{ and } H \text{ are Markov equivalent.} \tag{8.16}$$

This requirement is quite natural from the statistical point of view. Indeed, if acyclic directed graphs $G$ and $H$ are Markov equivalent then they represent the same statistical model and data 'generated' by $P \in \mathbb{M}_G = \mathbb{M}_H$ do not distinguish between $G$ and $H$. Thus, provided that the aim is to learn a statistical model, neither the respective quality criterion should distinguish between Markov equivalent acyclic directed graphs.

Remark 8.6 The concept of score equivalent criterion was pinpointed in § 4.2.5 of [13]. Note that most criteria for learning DAG models are score equivalent - see Observation 8.2 for examples. However, there are quality criteria which are not score equivalent. A well-known example, which is mentioned on p. 71 of [13], is a particular Bayesian criterion named the K2 metric in [55] after the algorithm in which it was first used [23]. A possible argument in favour of quality criteria which are not score equivalent is that if one is interested in a causal interpretation of acyclic directed graphs (see [117]) then a criterion of this kind allows one to distinguish between different causal interpretations. However, the point is that the difference in a causal interpretation cannot be identified on the basis of statistical data and therefore it is not a good idea to use such a criterion in the learning phase, which is based on data. $\triangle$

### 8.2.3 Decomposable criteria

Given a database $D \in \mathsf{DATA}(N,d) : x^1,\ldots,x^d,\ d \geq 1$ over $N$ and $\emptyset \neq A \subseteq N$, the database $D_A \in \mathsf{DATA}(A,d) : x^1_A,\ldots,x^d_A$ over $A$ will be called a *projection* of $D$ onto $A$. A quality criterion $\mathcal{Q}$ for learning DAG models will be called *decomposable* if there exists a collection of functions $q_{i|B} : \mathsf{DATA}(\{i\}\cup B, d) \to \mathbb{R}$, where $i \in N$, $B \subseteq N\setminus\{i\}$ and $d \in \mathbb{N}$, such that

$$\mathcal{Q}(G,D) \;=\; \sum_{i\in N} q_{i|pa_G(i)}(D_{i\cup pa_G(i)}) \tag{8.17}$$

for every $G \in \mathsf{DAGS}(N)$ and every $D \in \mathsf{DATA}(N,d)$. An important fact is that the functions $q_{i|B}$ do not depend on $G$; the graph $G$ is only represented in the right-hand side of (8.17) by the collection of sets $pa_G(i)$, $i \in N$. The criterion $\mathcal{Q}$ will be called *strongly decomposable* if, moreover, $q_{i|B}$ only depends on $D_{i\cup B}$ through the respective marginal contingency table $ct_{i\cup B}(D)$ (see p. 202 for this concept). More precisely, $\mathcal{Q}$ is strongly decomposable if a collection of functions $\bar{q}_{i|B} : \mathsf{CONT}(\{i\}\cup B, d) \to \mathbb{R}$, $i \in N$, $B \subseteq N\setminus\{i\}$, where $d \in \mathbb{N}$, exists such that

$$\mathcal{Q}(G,D) \;=\; \sum_{i\in N} \bar{q}_{i|pa_G(i)}(ct_{i\cup pa_G(i)}(D)) \tag{8.18}$$

for every $G \in \mathsf{DAGS}(N)$ and $D \in \mathsf{DATA}(N,d)$ (see Section 10.10.1).

Remark 8.7 The concept of decomposable criterion was pinpointed in [22]. Actually, all criteria for learning DAG models which are used in practice are decomposable owing to the way they are constructed - see [19]. This is caused by the intended application of the local search method (see p. 140). Indeed, if acyclic directed graphs $G$ and $H$ differ in the presence of one arrow then there exists $t \in N$ such that $pa_G(i) = pa_H(i)$ for every $i \in N \setminus t$ and by (8.17)

$$\mathcal{Q}(G,D) - \mathcal{Q}(H,D) \;=\; q_{t|pa_G(t)}(D_{t\cup pa_G(t)}) - q_{t|pa_H(t)}(D_{t\cup pa_H(t)})\,.$$

Thus, the decomposability of a criterion for learning DAG models is a requirement brought by the researchers in the area of artificial intelligence in order to make computations feasible. However, the definitions of the concept of decomposable criterion I found in the literature are slightly vague. The authors [19, 22] seem to repeat a sketchy phrase from [48]: a criterion is decomposable if it can be written as a product of 'measures', each of which is a function of only one node and its parents. What is not clear is how data affect the value of the criterion and what type of data is supposed to be given. Indeed, one can either consider data in the form of an ordered sequence of elements of the sample space or in the form of a contingency table. The distinction can be explained by means of the following simple example: if $y, z \in \mathsf{X}_N$, $y \neq z$ then the database $D : x^1 = y$, $x^2 = z$ of the length 2 differ from the database $D' : x^1 = z$, $x^2 = y$ if one accepts the interpretation of data in terms of an ordered sequence, but they coincide if one accepts the other interpretation! One can certainly imagine quality criteria whose values depend on the order of items in a database. In particular, it really matters what type of input data one has in mind because the other understanding of data confines the class of quality criteria.

Some explanation was only given by Chickering [22], who made an additional remark saying that data are implicit in the 'measures' and it is meant that the data on which a 'measure' depends are restricted to those columns which correspond to the respective node and its parents. This suggests that Chickering perhaps has in mind data in the form of an ordered sequence of items. On the other hand, in the formulas for various decomposable criteria [13, 19] data are represented in the form of a contingency table. Because two different understandings of data lead to two different concepts of decomposable quality criterion, I decided to include both versions and to distinguish between them by means of terminology: the concept of strong decomposability corresponds to the situation when data are supposed in the form of a contingency table. The distinction is also reflected in the subsequent concepts and results. $\triangle$

### 8.2.4 Regular criteria

Two traditional requirements on a quality criterion can be joined in the following concept. A quality criterion for learning DAG models will be called *regular* if there exists a collection of functions $\boldsymbol{t}_A : \mathsf{DATA}(A,d) \to \mathbb{R}$, $\emptyset \neq A \subseteq N$ and a constant depending on the sample space and $d$, by convention denoted by $\boldsymbol{t}_\emptyset(D_\emptyset)$, such that

$$\mathcal{Q}(G,D) \;=\; \sum_{i\in N} \boldsymbol{t}_{i\cup pa_G(i)}(D_{i\cup pa_G(i)}) - \boldsymbol{t}_{pa_G(i)}(D_{pa_G(i)}) \tag{8.19}$$

for every $G \in \mathsf{DAGS}(N)$ and every $D \in \mathsf{DATA}(N,d)$.

A criterion $\mathcal{Q}$ will be called *strongly regular* if $\boldsymbol{t}_A$ depend on $D_A$ through $ct_A(D)$, that is, if there exists a collection of functions $\bar{\boldsymbol{t}}_A : \mathsf{CONT}(A,d) \to \mathbb{R}$, $\emptyset \neq A \subseteq N$ and a constant $\bar{\boldsymbol{t}}_\emptyset(-)$ depending on the sample space and $d$ such that

$$\mathcal{Q}(G,D) \;=\; \sum_{i\in N} \bar{\boldsymbol{t}}_{i\cup pa_G(i)}(ct_{i\cup pa_G(i)}(D)) - \bar{\boldsymbol{t}}_{pa_G(i)}(ct_{pa_G(i)}(D)) \tag{8.20}$$

for every $G \in \mathsf{DAGS}(N)$ and every $D \in \mathsf{DATA}(N,d)$.

Observe that it follows immediately from this definition that a linear combination of (strongly) regular criteria is again a (strongly) regular criterion.

LEMMA 8.3 A quality criterion $\mathcal{Q}$ for learning DAG models is regular iff it is decomposable and score equivalent. Moreover, it is strongly regular iff it is strongly decomposable and score equivalent.

**Proof:** If $\mathcal{Q}$ is regular then put

$$q_{i|B}(D) \;=\; \boldsymbol{t}_{i\cup B}(D) - \boldsymbol{t}_B(D_B) \qquad \text{for } i \in N,\ B \subseteq N \setminus \{i\},\ D \in \mathsf{DATA}(\{i\} \cup B, d), \tag{8.21}$$

and observe that (8.19) implies (8.17). Transformational characterization of equivalent acyclic directed graphs (see p. 48) can simplify the proof of score equivalence of $\mathcal{Q}$: it suffices to verify for every $D \in \mathsf{DATA}(N,d)$ that $\mathcal{Q}(G,D) = \mathcal{Q}(H,D)$ whenever $H$ is obtained from $G \in \mathsf{DAGS}(N)$ by a legal reversal of an arrow $a \to b$ in $G$. The proof of this fact is analogous to the proof of Consequence 7.1: owing to (8.19) and the fact $pa_G(i) = pa_H(i)$ for $i \in N \setminus \{a, b\}$ one needs to evidence

$$\sum_{i \in \{a,b\}} \boldsymbol{t}_{i \cup pa_G(i)}(D_{i \cup pa_G(i)}) - \boldsymbol{t}_{pa_G(i)}(D_{pa_G(i)}) \;=\; \sum_{i \in \{a,b\}} \boldsymbol{t}_{i \cup pa_H(i)}(D_{i \cup pa_H(i)}) - \boldsymbol{t}_{pa_H(i)}(D_{pa_H(i)})\,.$$

This causes no problems since there exists $C \subseteq N$ such that $pa_G(a) = pa_H(b) = C$, $pa_G(b) = C \cup a$ and $pa_H(a) = C \cup b$.

To verify the converse implication, suppose that $\mathcal{Q}$ is decomposable and score equivalent. The first observation is that $\forall\, C \subseteq N$, $\forall\, a, b \in N \setminus C$, $a \neq b$ and $\forall\, D \in \mathsf{DATA}(N, d)$

$$q_{a|bC}(D_{abC}) + q_{b|C}(D_{bC}) \;=\; q_{b|aC}(D_{abC}) + q_{a|C}(D_{aC})\,. \tag{8.22}$$

Indeed, let us construct $G, H \in \mathsf{DAGS}(N)$ as follows. Both graphs have all arrows from $C$ to $\{a, b\}$. Moreover, $b \to a$ is an arrow in $G$ while $a \to b$ is an arrow in $H$. Clearly, $G$ and $H$ are equivalent acyclic directed graphs over $N$ and $\forall\, i \in N \setminus \{a, b\}$ $\;pa_G(i) = pa_H(i)$. It suffices to substitute (8.17) into (8.16); after cancellation of the terms which correspond to $i \in N \setminus \{a, b\}$, (8.22) is obtained.

The required system of functions $\boldsymbol{t}_A$, $A \subseteq N$ can be constructed recursively. First, put $\boldsymbol{t}_\emptyset(D_\emptyset) = 0$. For $\emptyset \neq A \subseteq N$ define $\boldsymbol{t}_A(D_A)$ on the basis of $\boldsymbol{t}_B(D_B)$ with $B \subset A$. Choose $a \in A$ and put

$$\boldsymbol{t}_A(D_A) \;=\; q_{a|A\setminus a}(D_A) + \boldsymbol{t}_{A\setminus a}(D_{A\setminus a}) \qquad \text{for } D \in \mathsf{DATA}(N, d)\,. \tag{8.23}$$

However, one has to show that this definition does not depend on the choice of $a \in A$, in other words, that (8.23) holds for any other $b \in A$ in place of $a$. This can be proved by induction on $|A|$. It is trivial if $|A| = 1$. If $|A| \geq 2$ observe that $\forall\, D \in \mathsf{DATA}(N, d)$, $\forall\, a, b \in A$, $a \neq b$

$$q_{a|A\setminus a}(D_A) + \boldsymbol{t}_{A\setminus a}(D_{A\setminus a}) \;=\; q_{b|A\setminus b}(D_A) + \boldsymbol{t}_{A\setminus b}(D_{A\setminus b})\,.$$

Indeed, put $C = A \setminus \{a, b\}$ in (8.22) and add the term $\boldsymbol{t}_{A\setminus ab}(D_{A\setminus ab})$ to both sides of the equality

$$q_{a|A\setminus a}(D_A) + \{\, q_{b|A\setminus ab}(D_{A\setminus a}) + \boldsymbol{t}_{A\setminus ab}(D_{A\setminus ab})\,\} \;=\; q_{b|A\setminus b}(D_A) + \{\, q_{a|A\setminus ab}(D_{A\setminus b}) + \boldsymbol{t}_{A\setminus ab}(D_{A\setminus ab})\,\}\,.$$

By the induction hypothesis, the expressions in braces are $\boldsymbol{t}_{A\setminus a}(D_{A\setminus a})$ and $\boldsymbol{t}_{A\setminus b}(D_{A\setminus b})$, which yields the desired equality.

The validity of (8.23) for every $A \subseteq N$ implies (8.21): put $i = a$ and $B = A \setminus a$. This in combination with (8.17) gives (8.19).

The proof of the claim concerning the strongly regular criterion is omitted. It is analogous: one simply writes $ct_A(D)$ instead of $D_A$, $\bar{q}$ instead of $q$ and $\bar{\boldsymbol{t}}$ instead of $\boldsymbol{t}$. □

The functions $\boldsymbol{t}_A$, $A \subseteq N$ inducing a regular criterion are not uniquely determined. To characterize those collections of functions which induce the same quality criterion, the following lemma characterizing special modular functions (see p. 81) is needed.

Lemma 8.4 Let $\mathcal{L}_*(N)$ denote the set of modular functions $l$ such that $l(N) = l(\emptyset)$. Then

$$\mathcal{L}_*(N) \;=\; \{l \in \mathbb{R}^{\mathcal{P}(N)} :\; \langle l, \delta_N - \delta_\emptyset \rangle = 0 \text{ and } \forall\, G \in \mathsf{DAGS}(N)\ \langle l, u_G \rangle = 0\}\,. \tag{8.24}$$

Moreover, $\mathcal{L}_*(N)$ is a linear subspace of $\mathbb{R}^{\mathcal{P}(N)}$ of dimension $|N|$ and

$$l \in \mathcal{L}_*(N) \;\Leftrightarrow\; l = w_\emptyset \cdot m^{\emptyset\uparrow} + \sum_{i \in N} w_i \cdot m^{\{i\}\uparrow} \quad \text{where } w_\emptyset, w_i \in \mathbb{R},\ \sum_{i \in N} w_i = 0\,. \tag{8.25}$$

**Proof:** To show (8.24), it suffices to show that $l \in \mathbb{R}^{\mathcal{P}(N)}$ is modular iff $\langle l, u_G \rangle = 0$ for every $G \in \mathsf{DAGS}(N)$.
Indeed, by Observation 5.1 $l \in \mathcal{L}(N)$ iff $\langle l, u \rangle = 0$ for every $u \in \mathcal{E}(N)$, respectively for every $u \in \mathcal{S}(N)$. However, by Remark 7.3 and Lemma 7.1, $\mathcal{E}(N) \subseteq \{u_G;\, G \in \mathsf{DAGS}(N)\} \subseteq \mathcal{S}(N)$.
By Lemma 5.2, $\mathcal{L}(N)$ is a linear subspace of dimension $|N|+1$. Thus, $\mathcal{L}_*(N) \subset \mathcal{L}(N)$ implies that $\mathcal{L}_*(N)$ is a linear subspace whose dimension does not exceed $|N|$. To show that its dimension is $|N|$ it suffices to construct a linearly independent subset of $\mathcal{L}_*(N)$ of cardinality $|N|$. To this end, fix a total ordering $b_1, \ldots, b_{|N|}$ of $N$ and put

$$l_1 = m^{\emptyset\uparrow}, \quad l_i = m^{\{b_{i-1}\}\uparrow} - m^{\{b_i\}\uparrow}, \quad \text{for } i = 2, \ldots, |N|\,. \tag{8.26}$$

By Lemma 5.2, $l_1, \ldots, l_{|N|} \in \mathcal{L}_*(N)$. Observe that they form a linearly independent set. Indeed, suppose that for some $\alpha_i \in \mathbb{R}$ one has $\sum_{i=1}^{|N|} \alpha_i \cdot l_i(A) = 0$ for any $A \subseteq N$. Substitute $A = \{b_i\}$ for $i = 1, \ldots, |N|$ and obtain

$$0 = \alpha_1 + \alpha_2, \quad 0 = \alpha_1 - \alpha_i + \alpha_{i+1} \ \text{ for } i = 2, \ldots, |N|-1 \quad \text{and} \quad 0 = \alpha_1 - \alpha_{|N|}\,.$$

Sum these equalities to get $0 = |N| \cdot \alpha_1$. Then substitute $\alpha_1 = 0$ into the equalities and show by induction that $\alpha_i = 0$ for $i = 2, \ldots, |N|$. Thus, $\alpha_1 = \ldots = \alpha_{|N|} = 0$ which concludes the proof of the fact that $l_1, \ldots, l_{|N|}$ are linearly independent.
The above observation means that $l_1, \ldots, l_{|N|}$ is a linear basis of $\mathcal{L}_*(N)$. Therefore, $l \in \mathcal{L}_*(N)$ iff $l = \sum_{i=1}^{|N|} \alpha_i \cdot l_i$ for some $\alpha_i \in \mathbb{R}$. Substitute (8.26) in this formula and get (8.25) where

$$w_\emptyset = \alpha_1, \ w_{\{b_1\}} = \alpha_2, \ w_{\{b_i\}} = -\alpha_i + \alpha_{i+1} \ \text{ for } i = 2, \ldots, |N|-1 \ \text{ and } w_{\{b_{|N|}\}} = -\alpha_{|N|}\,.$$

Because the above correspondence between $w$'s and $\alpha$'s is a one-to-one correspondence, a general expression for $l \in \mathcal{L}_*(N)$ is given by (8.25). □

Consequence 8.3 Let $\mathcal{Q}$ be a regular quality criterion for learning DAG models given by (8.19) for some $\boldsymbol{t}_A$, $A \subseteq N$. Then a collection of functions $\tilde{\boldsymbol{t}}_A : \mathsf{DATA}(A, d) \to \mathbb{R}$, $A \subseteq N$ (where $\tilde{\boldsymbol{t}}_\emptyset$ is a constant depending on $d$ and the sample space) defines $\mathcal{Q}$ by (8.19) iff there exist constants $w_\emptyset(d)$ and $w_i(d)$, $i \in N$ such that $\sum_{i \in N} w_i(d) = 0$ and

$$\tilde{\boldsymbol{t}}_A(D_A) \;=\; \boldsymbol{t}_A(D_A) + w_\emptyset(d) + \sum_{i \in A} w_i(d) \quad \text{for } A \subseteq N,\ D \in \mathsf{DATA}(N, d)\,. \tag{8.27}$$

**Proof:** Let us fix the sample space and the length $d$ of data. Then the collection of functions $\boldsymbol{t}_A$, $A \subseteq N$ can be interpreted as a real function on $\mathsf{DATA}(N, d) \times \mathcal{P}(N)$ which assigns the value $\boldsymbol{t}_A(D_A) \equiv t_D(A)$ to $D \in \mathsf{DATA}(N, d)$ and $A \subseteq N$. Moreover, this function, for every $\emptyset \neq A \subseteq N$, depends on $D_A$ and it is constant on $\mathsf{DATA}(N, d)$ if $A = \emptyset$. The collection $\tilde{\boldsymbol{t}}_A$, $A \subseteq N$ can be represented in an analogous way. Now, by (8.19) and the definition of a standard imset (see p. 116), derive that $\boldsymbol{t}_A$, $A \subseteq N$ and $\tilde{\boldsymbol{t}}_A$, $A \subseteq N$ both induce $\mathcal{Q}$ iff

$$\forall\, D \in \mathsf{DATA}(N, d)\ \forall\, G \in \mathsf{DAGS}(N) \quad \langle t_D - \tilde{t}_D, \delta_N - \delta_\emptyset - u_G \rangle = 0\,.$$

Observe that this condition is equivalent to the requirement that $l_D \equiv t_D - \tilde{t}_D \in \mathcal{L}_*(N)$ for every $D \in \mathsf{DATA}(N, d)$.

Indeed, for a fixed $D \in \mathsf{DATA}(N,d)$, choose $G \in \mathsf{DAGS}(N)$ with $u_G = 0$ and observe $\langle l_D, \delta_N - \delta_\emptyset \rangle = 0$. Hence $\langle l_D, u_G \rangle = 0$ for every $G \in \mathsf{DAGS}(N)$. These two conditions are equivalent to $l_D \in \mathcal{L}_*(N)$ by (8.24) in Lemma 8.4.
Now, let us apply (8.25) to every $l_D$, $D \in \mathsf{DATA}(N,d)$ to see that the requirement is equivalent to the existence of functions $w_\emptyset, w_i : \mathsf{DATA}(N,d) \to \mathbb{R}$, $i \in N$ such that $\sum_{i \in N} w_i(D) = 0$ and

$$l_D(A) = w_\emptyset(D) \cdot m^{\emptyset\uparrow}(A) + \sum_{i \in N} w_i(D) \cdot m^{\{i\}\uparrow}(A) \qquad \text{for } D \in \mathsf{DATA}(N,d),\ A \subseteq N. \quad (8.28)$$

The point is that these functions have to be constant.
Indeed, the substitution $A = \emptyset$ in (8.28) gives $w_\emptyset(D) = l_D(\emptyset) = \boldsymbol{t}_\emptyset(D_\emptyset) - \tilde{\boldsymbol{t}}_\emptyset(D_\emptyset)$ which is a constant. Analogously, by putting $A = \{i\}$ for $i \in N$ we derive that $w_i$ depends on $D_i$. Suppose for contradiction that $w_i$ is not constant for some $i \in N$. Then databases $D^1 : x^1, \ldots, x^d$ and $D^2 : y^1, \ldots, y^d$ exist such that $w_i(D^1) \neq w_i(D^2)$. Construct a database $D^3 : z^1, \ldots, z^d$ by putting $z^\ell = [x^\ell_{N\setminus i}, y^\ell_i]$ for $\ell = 1, \ldots d$. Thus, $D^3_{N\setminus i} = D^1_{N\setminus i}$ and $D^3_i = D^2_i$. As $w_{i'}$, $i' \in N$ depend on $D_{i'}$ one has $\sum_{i' \in N} w_{i'}(D^3) - \sum_{i' \in N} w_{i'}(D^1) = w_i(D^2) - w_i(D^1) \neq 0$ which contradicts the assumption $\sum_{i \in N} w_i(D) = 0$ for any $D \in \mathsf{DATA}(N,d)$.
On the other hand, the validity of (8.28) for every $D \in \mathsf{DATA}(N,d)$ and constant functions $w_\emptyset$ and $w_i$, $i \in N$ is nothing but (8.27). $\square$

Remark 8.8 The class $\mathcal{L}_*(N)$ induces an equivalence relation $t - \tilde{t} \in \mathcal{L}_*(N)$ for $t, \tilde{t} \in \mathbb{R}^{\mathcal{P}(N)}$. One can introduce a reasonable 'standard' representative of every equivalence class then, namely the only $t_*$ within the class such that $t_*(A) = 0$ for $A \subseteq N$ with $|A| = 1$. However, the idea is not applicable in the situation where one is looking for a 'standardized' collection of functions $\boldsymbol{t}_A$, $A \subseteq N$ inducing a regular criterion. The reason is that $\boldsymbol{t}_i$ need not be constant for $i \in N$ while the available 'standardization' operation (8.27) deals with constants. Thus, one can achieve by 'standardization' $\boldsymbol{t}_i(D_i) = 0$ for every $i \in N$ and any chosen database $D \in \mathsf{DATA}(N,d)$ but then another database $D'$ may exist for which this condition is not valid. $\triangle$

The formulas derived in Section 8.2.1 allow one to verify that basic information criteria are strongly regular.

Observation 8.1 The maximized log-likelihood criterion (MLL) is a strongly regular criterion for learning DAG models.

**Proof:** Suppose the situation described by Convention 3. Put

$$\bar{\boldsymbol{t}}_A(\boldsymbol{d}) = \begin{cases} \sum_{x \in \mathsf{X}_A} \boldsymbol{d}(x) \cdot \ln \boldsymbol{d}(x) & \text{if } \emptyset \neq A \subseteq N,\ \boldsymbol{d} \in \mathsf{CONT}(A,d), \\ d \cdot \ln d & \text{if } A = \emptyset, \end{cases} \qquad (8.29)$$

where the convention $0 \cdot \ln \frac{0}{*}$ is accepted. By (8.9), the formula $d_{ij} = \sum_{k=1}^{r(i)} d_{ijk}$ and Convention 3 we can write:

$$\mathsf{MLL}\,(G,D) \;=\; \sum_{i \in N} \sum_{j=1}^{q(i,G)} \sum_{k=1}^{r(i)} d_{ijk} \cdot \ln \frac{d_{ijk}}{d_{ij}} \;=$$

$$= \sum_{i\in N}\sum_{j=1}^{q(i,G)}\sum_{k=1}^{r(i)} d_{ijk}\cdot \ln d_{ijk} - \sum_{i\in N}\sum_{j=1}^{q(i,G)} \ln d_{ij}\cdot \sum_{k=1}^{r(i)} d_{ijk} =$$

$$= \sum_{i\in N}\{ \underbrace{\sum_{j=1}^{q(i,G)}\sum_{k=1}^{r(i)} d_{ijk}\cdot \ln d_{ijk}}_{\bar{\boldsymbol{t}}_{i\cup pa_G(i)}(ct_{i\cup pa_G(i)}(D))} - \underbrace{\sum_{j=1}^{q(i,G)} d_{ij}\cdot \ln d_{ij}}_{\bar{\boldsymbol{t}}_{pa_G(i)}(ct_{pa_G(i)}(D))} \}.$$

for $G \in \mathsf{DAGS}(N)$ and $D \in \mathsf{DATA}(N,d)$. $\square$

The mapping which assigns the effective dimension of $\mathbb{M}_G$ to every $G \in \mathsf{DAGS}(N)$ can be viewed as a special quality criterion for learning DAG models, namely a criterion which does not depend on data.

Observation 8.2 The effective dimension (DIM) is a strongly regular criterion for learning DAG models. Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) are strongly regular criteria as well. In particular, they are score equivalent and strongly decomposable.

**Proof:** Supposing Convention 3 put

$$\bar{\boldsymbol{t}}'_A(\boldsymbol{d}) = \begin{cases} \prod_{l\in A} r(l) & \text{if } \emptyset \neq A \subseteq N, \quad \text{for every } \boldsymbol{d} \in \mathsf{CONT}(A,d), \\ 1 & \text{if } A = \emptyset. \end{cases} \tag{8.30}$$

By (8.11) and Convention 3 for every $G \in \mathsf{DAGS}(N)$ and $D \in \mathsf{DATA}(N,d)$ we can now write:

$$\mathsf{DIM}\,(G,D) = \sum_{i\in N}(r(i)-1)\cdot q(i,G) = \sum_{i\in N} r(i)\cdot q(i,G) - \sum_{i\in N} q(i,G) =$$

$$= \sum_{i\in N}\{ \underbrace{\prod_{l\in\{i\}\cup pa_G(i)} r(l)}_{\bar{\boldsymbol{t}}'_{i\cup pa_G(i)}(ct_{i\cup pa_G(i)}(D))} - \underbrace{\prod_{l\in pa_G(i)} r(l)}_{\bar{\boldsymbol{t}}'_{pa_G(i)}(ct_{pa_G(i)}(D))} \}.$$

Thus, the DIM criterion is strongly regular. This fact, together with Obervation 8.1 and the definitions of the AIC criterion and the BIC criterion, implies the same conclusion for these criteria. Finally, use Lemma 8.3. $\square$

Remark 8.9 In my view, the logarithm of the marginal likelihood (LML) mentioned in Remark 8.5 can also be shown to be strongly regular if relevant assumptions on the priors $\pi_G$, $G \in \mathsf{DAGS}(N)$ are accepted. More exactly, I have in mind the assumptions explicated in [28], namely a technical assumption that $\pi_G$ are Dirichlet measures and the assumption that the priors are compatible (see § 6 of [28]). Actually, as mentioned on p. 54 of [19], there are examples of Bayesian criteria which are both decomposable and score equivalent. Therefore, they are regular by Lemma 8.3. $\triangle$

## 8.3 Inclusion neighbourhood

The binary relation $\mathcal{M}_K \subseteq \mathcal{M}_L$ for $K, L \in \mathsf{DAGS}(N)$ defines the *inclusion quasiordering* on the set of acyclic directed graphs over $N$. It can also be viewed as a partial ordering on the set of Markov equivalence classes of acyclic directed graphs over $N$, respectively on the set of DAG models over $N$. In this chapter, every $G \in \mathsf{DAGS}(N)$ is also interpreted as a statistical model $\mathbb{M}_G$ on a (fixed) discrete sample space $\mathsf{X}_N = \prod_{i \in N} \mathsf{X}_i$ (see p. 142). Clearly, $\mathcal{M}_K \subseteq \mathcal{M}_L$ implies $\mathbb{M}_K \supseteq \mathbb{M}_L$ and the converse is true if $|\mathsf{X}_i| \geq 2$ for every $i \in N$ because of the existence of a perfectly Markovian measure on $\mathsf{X}_N$ for every $G \in \mathsf{DAGS}(N)$ [88]. In particular, the *strict inclusion* of formal independence models $\mathcal{M}_K \subset \mathcal{M}_L$ for $K, L \in \mathsf{DAGS}(N)$ is equivalent to strict inclusion $\mathbb{M}_K \supset \mathbb{M}_K$ of statistical models then.

We say that $K, L \in \mathsf{DAGS}(N)$ are *inclusion neighbours* and write $\mathcal{M}_K \sqsubset \mathcal{M}_L$ if $\mathcal{M}_K \subset \mathcal{M}_L$ and there is no $G \in \mathsf{DAGS}(N)$ such that $\mathcal{M}_K \subset \mathcal{M}_G \subset \mathcal{M}_L$. More precisely, we say then that $\mathcal{M}_L$ is the *upper neighbour* of $\mathcal{M}_K$ or, dually, that $\mathcal{M}_K$ is the *lower neighbour* of $\mathcal{M}_L$. The *inclusion neighbourhood* of $\mathcal{M}_G, G \in \mathsf{DAGS}(N)$ consists of the union of the collection of upper neighbours and the collection of lower neighbours.

Remark 8.10 The inclusion ordering can be understood in two ways: either in the sense $\mathcal{M}_K \subseteq \mathcal{M}_L$ or in the sense $\mathbb{M}_K \supseteq \mathbb{M}_L$. Thus, it is a matter of taste how this asymmetric binary relation of graphs $K$ and $L$ is reflected in the terminology, notation and pictures. Whittaker [148] had in mind the interpretation of graphs in terms of statistical models. Thus, in his view, if $\mathbb{M}_K \supseteq \mathbb{M}_L$ then $K$ represents a bigger statistical model than $L$. Therefore, he called the saturated model, which is represented by a complete graph, the 'maximal' graphical model (see p. 185 of [148]) and always placed the pictorial representatives of this model at the top of his lattice diagrams. Chickering [22] used the formula $L \leq K$ to denote $\mathcal{M}_K \subseteq \mathcal{M}_L$. It seems that he was mainly led by graphical pictorial representation. Indeed, $\mathcal{M}_K \subseteq \mathcal{M}_L$ implies that $K$ contains at least those edges which are present in $L$. Similarly, in [55, 19], graphs with maximal numbers of edges are at the top of the respective Hasse diagrams.

However, CI interpretation is the main issue in this monograph. Therefore I prefer the interpretation in terms of the formal independence model, that is, $\mathcal{M}_K \subseteq \mathcal{M}_L$. One reason for my preference is that this point of view admits a wider perspective: graphical models can be interpreted as the elements of the lattice of CI structures. This approach is naturally reflected in terminology and pictures. Therefore, if $\mathcal{M}_K \sqsubset \mathcal{M}_L$, then $\mathcal{M}_L$ is called an upper neighbour of $\mathcal{M}_K$ and the pictorial representative of $\mathcal{M}_L$ is put above the pictorial representative of $\mathcal{M}_K$ in all diagrams in this book. In particular, the saturated model, which represents the largest class of probability measures but empty CI information, is always placed at the bottom of a Hasse diagram. Note that in order to avoid confusion I intentionally do not use a notation like $L \leq K$. $\triangle$

The concept of an inclusion neighbourhood is certainly important from the theoretical point of view. Nevertheless, there are practical reasons for which it deserves serious attention. These reasons are strongly related to the local search method mentioned on p. 140. It is claimed on p. 66 of [19] that the *inclusion boundary condition* of [55] is a necessary condition for a neighbourhood structure to avoid an unpleasant situation in which a local search procedure finds a local maximum of a quality criterion which is not the global maximum. This condition requires, for every $G \in \mathsf{DAGS}(N)$, that the inclusion neighbourhood of $\mathcal{M}_G$ should be covered by $\bigcup\{\mathcal{M}_H; H \in \mathit{nei}(G)\}$ where $\mathit{nei}(G)$ is the class of neighbouring graphs for $G$.

Actually, the significance of this condition was implicitly recognized already in [89] in connection with a particular local search procedure named *greedy equivalence search* (GES) algorithm. The goal of the GES algorithm is learning DAG models. Every DAG model over $N$ can be represented graphically by any member of the respective equivalence class of acyclic directed graphs over $N$. Therefore, the quality criteria used in this procedure are supposed to be both decomposable and score equivalent. The states of the procedure are in fact equivalence classes of graphs, although the procedure formally works with individual elements of $\mathsf{DAGS}(N)$. The algorithm has two phases. In the first phase edges are inserted and in the second phase edges are removed. Thus, in each phase a different part of the overall neighbourhood structure of a 'current' graph is considered.

**(i)** The first phase starts with an acyclic directed graph $G$ over $N$ which has no arrows. In this phase, the set of neighbouring graphs of a current $G \in \mathsf{DAGS}(N)$ consists of $K \in \mathsf{DAGS}(N)$ such that there exists $L \in \mathsf{DAGS}(N)$ which is equivalent to $G$ and $K$ is obtained from $L$ by one arrow addition. Provided there exists a neighbouring graph $K$ which has a strictly positive increase $\mathcal{Q}(K,D) - \mathcal{Q}(G,D)$, $D \in \mathsf{DATA}(N,d)$ of the quality criterion value, one chooses $K$ with the maximum increase and moves from $G$ to $K$, which becomes a new current graph. The procedure is repeated until the quality criterion value in the current graph cannot be increased in this way.

**(ii)** The second phase starts with the result of the first phase. It employs the complementary part of the overall neighbourhood structure. In the second phase, the set of graphs neighbouring the current $G \in \mathsf{DAGS}(N)$ consists of $L \in \mathsf{DAGS}(N)$ such that there exists $K \in \mathsf{DAGS}(N)$ which is equivalent to $G$, and $L$ is obtained from $K$ by the removal of an arrow. Again, one is trying to maximize the increase $\mathcal{Q}(L,D) - \mathcal{Q}(G,D)$ of the quality criterion value and the procedure stops when the value of $\mathcal{Q}$ in the current graph cannot be increased in this way. The output of the GES algorithm is the equivalence class of acyclic directed graphs over $N$ which contains the last current graph $G \in \mathsf{DAGS}(N)$.

Meek [89] formulated a conjecture about a transformational graphical characterization of inclusion quasiordering for acyclic directed graphs (see Lemma 8.5 below). As reported in [22], Meek also showed that if his conjecture is true and many other assumptions are fulfilled, then the GES algorithm should find the "optimal" solution. That means, one should get that equivalence class of acyclic directed graphs with respect to which the probability measure which "generates" data is perfectly Markovian. Note that his assumptions involve the data faithfulness assumption (see p. 135, several other technical statistical assumptions and the assumption that the length of the database $D$ approaches to infinity. The validity of the conjecture is needed to show that the GES algorithm cannot terminate with a model that is not the optimal solution – for details see the proof of Lemma 4 in [22]. The conjecture about a transformational characterization of inclusion quasiordering has been recently confirmed in [22] with the help of a method used in [56] to verify the conjecture in a special case. The characterization is based on the operation of legal arrow reversal (see p. 48) and the operation of *legal arrow addition* by which an arrow addition into an acyclic directed graph is meant such that the resulting directed graph remains acyclic. However, I prefer a simpler formulation of the result, which uses the inverse operation. By an *arrow removal* the following change of an acyclic directed graph $K$ over $N$ into a directed graph $L$ over $N$ is understood: one simply removes an arrow $a \to b$ in $K$. Evidently, $L$ is again an acyclic directed graph. Moreover, one has $\mathcal{M}_K \subset \mathcal{M}_L$ (which follows from the following lemma).

LEMMA 8.5 Suppose that $K, L \in \mathsf{DAGS}(N)$. Then $\mathcal{M}_K \subseteq \mathcal{M}_L$ iff there exists a sequence $K = G_1, \ldots, G_n = L$, $n \geq 1$ of acyclic directed graphs over $N$ such that for every $i = 1, \ldots, n-1$

- either $G_{i+1}$ is obtained from $G_i$ by a legal arrow reversal, or
- $G_{i+1}$ is obtained from $G_i$ by an arrow removal.

In particular, $\mathcal{M}_K \subset \mathcal{M}_L$ if a sequence of graphs of this type exists such that $n \geq 2$ and the arrow removal operation is applied at least once in the sequence.
Finally, $\mathcal{M}_K \sqsubset \mathcal{M}_L$ if there exists $K', L' \in \mathsf{DAGS}(N)$ such that $K'$ is equivalent to $K$, $L'$ is equivalent to $L$, and $L'$ is obtained from $K'$ by an arrow removal.

**Proof:** The first claim of the lemma is proved in [22]. However, it is formulated in a "converse" order; namely $\mathcal{M}_K \subseteq \mathcal{M}_L$ if there exists a sequence $L = H_1, \ldots, H_n = K$, $n \geq 1$ such that, for every $j = 1, \ldots, n-1$, $H_{j+1}$ is obtained from $H_j$ either by a legal arrow reversal or by a legal arrow addition.

The second claim of the lemma can be verified as follows: the necessity of the condition follows from the first claim and the transformational characterization of equivalence of acyclic directed graphs mentioned on p. 48. The sufficiency follows from the equivalence characterization and the observation that if $L$ is obtained from $K \in \mathsf{DAGS}(N)$ by the removal of an edge then $\mathcal{M}_K \subset \mathcal{M}_L$.
Indeed, $\mathcal{M}_K \subseteq \mathcal{M}_L$ follows from the definition of the induced DAG model (p. 47): every route that is active in $L$ is an active route in $K$ which means that every route blocked in $K$ has to be blocked in $L$. The fact $\mathcal{M}_K \neq \mathcal{M}_L$ follows from the graphical characterization of Markov equivalence of acyclic directed graphs (p. 47): equivalent graphs have the same underlying graph.

The condition in the third claim is equivalent to the requirement that there exists a sequence $K = G_1, \ldots, G_n = L$, $n \geq 2$ of the considered type such that the arrow removal operation is applied exactly once (use the transformational characterization of equivalence). Thus, the third claim can be derived from the second one.
Indeed, if there are two arrow removals in the sequence then one has $\mathcal{M}_K \subset \mathcal{M}_G \subset \mathcal{M}_L$ for any graph $G$ in the sequence which is after the first removal and before the last removal. Conversely, if $\mathcal{M}_K \subset \mathcal{M}_G \subset \mathcal{M}_L$ for an acyclic directed graph $G$ then by the second claim of the lemma $|\mathcal{A}(K)| > |\mathcal{A}(G)| > |\mathcal{A}(L)|$, which implies that $K$ has at least two more edges than $L$. $\square$

Thus, a transformational characterization of inclusion quasiordering for acyclic directed graphs is available. However it does not seem to be very suitable for testing inclusion of DAG models which is the task to decide, given $K, L \in \mathsf{DAGS}(N)$, whether $\mathcal{M}_K \subseteq \mathcal{M}_L$ holds or not. Indeed, one has to try to construct a sequence of graphs between $K$ and $L$ and it is not clear how to do it. The question of whether a suitable direct graphical characterization of inclusion quasiordering for acyclic directed graphs exists remains open - see Theme 2. On the other hand, a special arithmetic criterion is proposed in the next section.

## 8.4 Standard imsets and learning

The method of structural imsets can be applied in the area of learning DAG models. It brings a certain simplifying perspective.

### 8.4.1 Inclusion neighbourhood characterization

Standard imsets for acyclic directed graphs (see p. 116) allow one to characterize the inclusion quasiordering in an arithmetic way.

OBSERVATION 8.3 Let $K, L \in \mathsf{DAGS}(N)$ be such that $L$ is obtained from $K$ by the removal of an arrow $a \to b$ in $K$. Then $u_L - u_K = u_{\langle a,b|C\rangle}$ where $C = pa_K(b) \setminus \{a\} = pa_L(b)$.

**Proof:** Since $pa_L(c) = pa_K(c)$ for $c \in N \setminus b$, $pa_L(b) = C$ and $pa_K(b) = aC$, by (7.3) we get $u_L - u_K = \{\delta_{pa_L(b)} - \delta_{bpa_L(b)}\} - \{\delta_{pa_K(b)} - \delta_{bpa_K(b)}\} = \delta_C - \delta_{bC} - \delta_{aC} + \delta_{abC}$. $\square$

LEMMA 8.6 Suppose $K, L \in \mathsf{DAGS}(N)$. Then $\mathcal{M}_K \subseteq \mathcal{M}_L$ ($\equiv$ $\mathbb{M}_K \supseteq \mathbb{M}_L$) iff $u_L - u_K$ is a combinatorial imset over $N$. Moreover, $\mathcal{M}_K \subset \mathcal{M}_L$ ($\equiv$ $\mathbb{M}_K \supset \mathbb{M}_L$) iff $u_L - u_K$ is a non-zero combinatorial imset.

**Proof:** If $\mathcal{M}_K \subseteq \mathcal{M}_L$ then apply Lemma 8.5 and consider the respective sequence $K = G_1, \ldots, G_n = L$, $n \geq 1$. If $G_{i+1}$ is obtained from $G_i$ by a legal arrow reversal then $G_{i+1}$ is Markov equivalent to $G_i$ and $u_{G_{i+1}} = u_{G_i}$ by Consequence 7.1. If $G_{i+1}$ is obtained from $G_i$ by an arrow removal then $u_{G_{i+1}} - u_{G_i} \in \mathcal{E}(N)$ by Observation 8.3. Thus $u_L - u_K = \sum_{i=1}^{n-1} u_{G_{i+1}} - u_{G_i}$ is a combinatorial imset. Conversely, if $u_L - u_K \in \mathcal{C}(N)$ then by Lemma 6.1 $u_L \rightharpoonup u_K$, that is, $\mathcal{M}_{u_K} \subseteq \mathcal{M}_{u_L}$. This means $\mathcal{M}_K \subseteq \mathcal{M}_L$ by Lemma 7.1.

Thus, the proved fact implies that $\mathcal{M}_K \subset \mathcal{M}_L$ iff $\{\, u_L - u_K \in \mathcal{C}(N)$ & $u_K - u_L \notin \mathcal{C}(N) \,\}$. However, it follows from Observation 4.4 for $w \in \mathcal{C}(N)$ that $-w \in \mathcal{C}(N) \Leftrightarrow w = 0$. Hence, the latter condition is that $u_L - u_K$ is a non-zero combinatorial imset. $\square$

Note that the above characterization gives a simple arithmetic criterion for testing inclusion of DAG models. Indeed, as mentioned in Remark 6.3, the question of whether an imset is a combinatorial imset is decidable.

CONSEQUENCE 8.4 Suppose $K, L \in \mathsf{DAGS}(N)$. Then $\mathcal{M}_K \sqsubset \mathcal{M}_L$ ($\equiv \mathbb{M}_K \sqsupset \mathbb{M}_L$) iff $u_L - u_K$ is an elementary imset.

**Proof:** To prove the necessity of the condition, use the third claim of Lemma 8.5: by Consequence 7.1 and Observation 8.3 we get $u_L - u_K = u_{L'} - u_{K'} \in \mathcal{E}(N)$. To prove the sufficiency, use Lemma 8.6 to derive $\mathcal{M}_K \subset \mathcal{M}_L$ and suppose for contradiction that $\mathcal{M}_K \subset \mathcal{M}_G \subset \mathcal{M}_L$ for some $G \in \mathsf{DAGS}(N)$. Then observe again by Lemma 8.6 that $u_L - u_K = (u_L - u_G) + (u_G - u_K)$ is the sum of two non-zero combinatorial imsets. Thus, $deg(u_L - u_K) \geq 2$ and $u_L - u_K \notin \mathcal{E}(N)$ by Consequence 4.2, which contradicts the assumption. $\square$

Of course, given a pair $K, L \in \mathsf{DAGS}(N)$ such that $\mathcal{M}_K \sqsubset \mathcal{M}_L$, the elementary imset $u_L - u_K$ is determined uniquely. It will be called the *differential imset* for $K$ and $L$.

REMARK 8.11 The concept of a differential imset brings an interesting interpretation of the edges in the Hasse diagram of the poset of DAG models over $N$. Every link of this diagram can be annotated by the respective elementary triplet and one can interpret the links of the diagram in terms of elementary conditional (in)dependence statements. An example of a diagram of this type for $N$ with $|N| = 3$ is given in Figure 8.1. Note that the DAG models are represented by standard imsets in the figure and that graphical representatives, namely the essential graphs which coincide with the largest chain graphs for $|N| \leq 3$ are attached for the reader's convenience.

The above-mentioned interpretation becomes relevant in connection with the local search method for learning DAG models (see p. 140). It was explained in Section 8.3 that, from both the theoretical and practical points of view, it is natural to consider the search space (of unique representatives) of DAG models endowed with the neighbourhood structure given by the concept of an inclusion neighbourhood. This search space is nothing but the Hasse diagram mentioned above. The point is that the 'moves' in the search space between neighbouring states have natural interpretation in terms of elementary conditional (in)dependence statements. This fact becomes even easier to see when one considers a particular regular criterion for learning DAG models - see Remark 8.13. Note, however, that this particular CI interpretation of moves between states of the search space in only possible in the context of DAG models - see Example 9.4.

### 8.4.2 Regular criteria and standard imsets

LEMMA 8.7 Let $\mathcal{Q}$ be a regular quality criterion for learning DAG models. Then there exist two statistics, namely, a function $s : \mathsf{DATA}(N,d) \rightarrow \mathbb{R}$ and a mapping $t : D \in \mathsf{DATA}(N,d) \mapsto t_D \in \mathbb{R}^{\mathcal{P}(N)}$, such that

- $\forall\, A \subseteq N\ |A| \leq 1 \quad t_D(A) = 0$ for every $D \in \mathsf{DATA}(N,d)$,
- $\forall\, A \subseteq N\ |A| \geq 2$ the coordinate mapping $D \mapsto t_D(A)$ depends on $D_A$,

and the formula

$$\mathcal{Q}(G,D) \;=\; s(D) - \langle t_D, u_G \rangle \tag{8.31}$$

holds for every $G \in \mathsf{DAGS}(N)$ and $D \in \mathsf{DATA}(N,d)$. The function $s$ and the mapping $t$ are uniquely detemined by these two requirements. Moreover, if $\mathcal{Q}$ is strongly regular, then

- $\forall\, A \subseteq N\ |A| \geq 2$ the mapping $D \mapsto t_D(A)$ depends on the contingency table $ct_A(D)$.

**Proof:** Supposing (8.19) holds for a collection $\boldsymbol{t}_A : \mathsf{DATA}(A,d) \rightarrow \mathbb{R}$, $A \subseteq N$, let us put

$$t'_D(A) \;=\; \boldsymbol{t}_A(D_A) \quad \text{for } D \in \mathsf{DATA}(N,d),\ A \subseteq N. \tag{8.32}$$

Thus, $t'_D \in \mathbb{R}^{\mathcal{P}(N)}$ for every data $D$. Introduce for every $D \in \mathsf{DATA}(N,d)$

$$\begin{aligned} s(D) &= t'_D(N) - t'_D(\emptyset) = \langle t'_D, \delta_N - \delta_\emptyset \rangle \text{ and} && (8.33)\\ t_D &= t'_D - t'_D(\emptyset)\cdot m^{\emptyset\uparrow} - \sum_{i\in N} \{\, t'_D(\{i\}) - t'_D(\emptyset)\,\} \cdot m^{\{i\}\uparrow} && (8.34) \end{aligned}$$

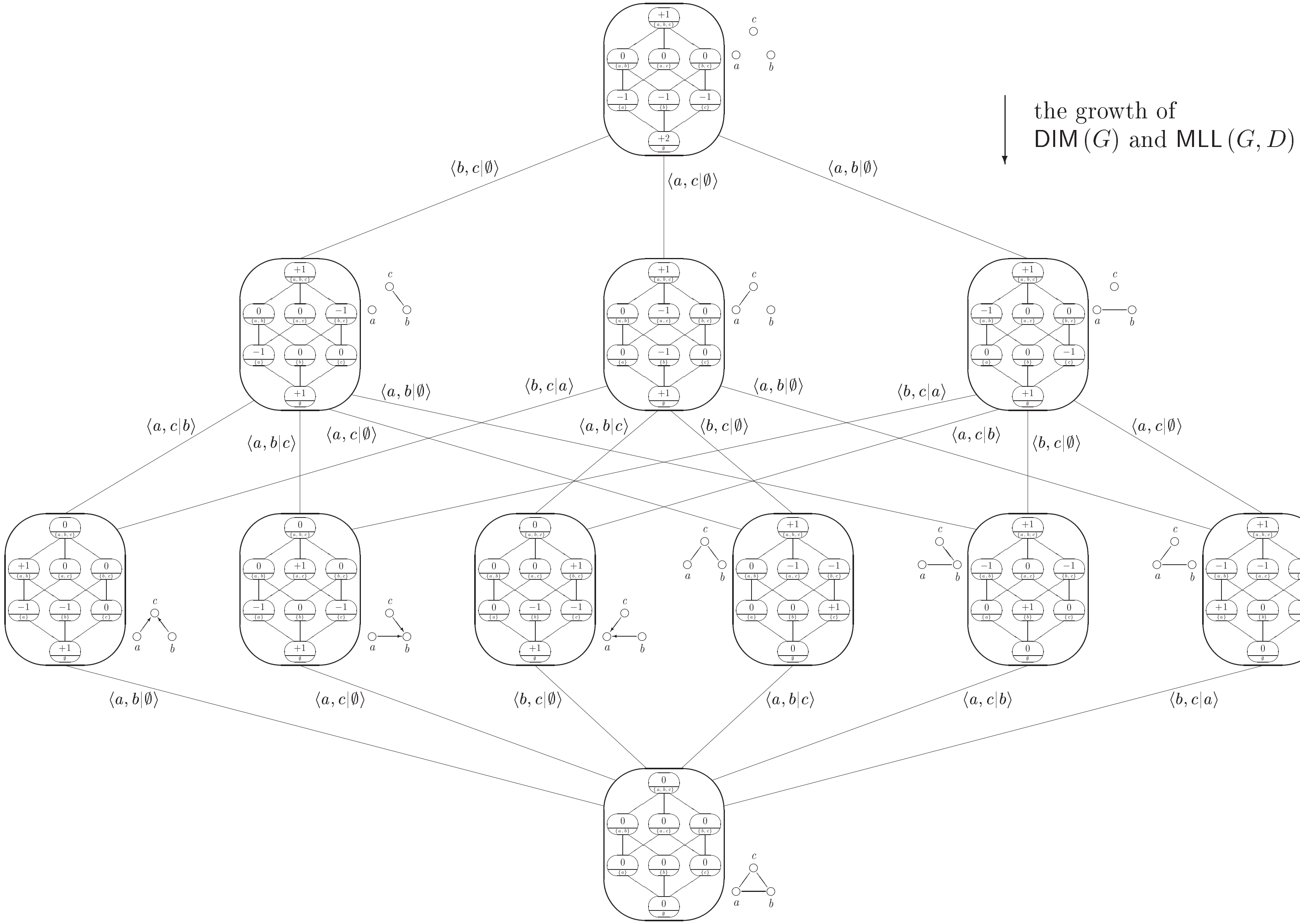


Figure 8.1: The search space of standard imsets over $N = \{a, b, c\}$ (rotated).

(see p. 41). This definition ensures $t_D(A) = 0$ for every $D$ and $|A| \leq 1$. Moreover, for every $D$ and $|A| \geq 2$ one has

$$t_D(A) \;=\; \boldsymbol{t}_A(D_A) - \sum_{i \in A} \boldsymbol{t}_{\{i\}}(D_i) + (|A| - 1) \cdot \boldsymbol{t}_\emptyset(D_\emptyset),$$

which implies that the mapping $D \mapsto t_D(A)$ depends on $D_A$. Observe that if $\mathcal{Q}$ is strongly regular then $t_D(A)$ depends on $ct_A(D)$ since $ct_B(D)$ is a function of $ct_A(D)$ if $B \subseteq A$.

By Lemma 5.2, $t_D - t'_D \in \mathcal{L}(N)$ for every $D$, which implies by Observation 5.1 that $\langle t_D - t'_D, u\rangle = 0$ for any $u \in \mathcal{S}(N)$. Hence, $\langle t_D, u_G\rangle = \langle t'_D, u_G\rangle$ for every $G \in \mathsf{DAGS}(N)$, $D \in \mathsf{DATA}(N, d)$. Thus, (8.19) takes the form

$$\mathcal{Q}(G, D) \;=\; \langle t'_D, \delta_N - \delta_\emptyset\rangle - \langle t'_D, u_G\rangle \;=\; s(D) - \langle t_D, u_G\rangle\,.$$

To evidence the uniqueness of $s$, take $H \in \mathsf{DAGS}(N)$ with $u_H = 0$ and observe $s(D) = \mathcal{Q}(H, D)$ for every $D$ by (8.31). To verify the uniqueness of $t$, suppose that another mapping $\tilde{t} : D \in \mathsf{DATA}(N, d) \mapsto \tilde{t}_D \in \mathbb{R}^{\mathcal{P}(N)}$ satisfies (8.31). Then $\langle t_D - \tilde{t}_D, u_G\rangle = 0$ for every $D \in \mathsf{DATA}(N, d)$ and $G \in \mathsf{DAGS}(N)$. By Remark 7.3, $\mathcal{E}(N) \subseteq \{u_G;\, G \in \mathsf{DAGS}(N)\}$ and therefore $t_D - \tilde{t}_D \in \mathcal{L}(N)$ for every $D$ (use Observation 5.1). As $t_D - \tilde{t}_D$ is $\ell$-standardized, $t_D - \tilde{t}_D \equiv 0$ by Remark 5.3. □

REMARK 8.12 The condition from Lemma 8.7 is necessary for the regularity of a quality criterion but it alone is not sufficient. To evidence this in the case of non-trivial sample spaces, that is, if $|\mathsf{X}_i| \geq 2$ for every $i \in N$, consider a statistic $s : \mathsf{DATA}(N, d) \to \mathbb{R}$ which cannot be written in the form of the sum $s(D) = \sum_{i \in N} s_i(D_i)$ where $s_i : \mathsf{DATA}(\{i\}, d) \to \mathbb{R}$, $i \in N$. The existence of $s$ follows easily from the dimension comparison of both linear spaces of functions. Then put $t_D = 0$ for any $D$ and define $\mathcal{Q}$ by (8.31): $\mathcal{Q}(G, D) = s(D)$ for any $G \in \mathsf{DAGS}(N)$, $D \in \mathsf{DATA}(N, d)$. However, $\mathcal{Q}$ is not regular because $s$ was constructed not to be decomposable (consider a graph $G \in \mathsf{DAGS}(N)$ which has no arrow).

Another warning is that even if $\mathcal{Q}$ is a regular criterion given by (8.31) the mapping $A \mapsto t_D(A)$ from Lemma 8.7 need not induce $\mathcal{Q}$ in the sense of (8.19). △

Given a particular quality criterion $\mathcal{Q}$ for learning DAG models, the symbol $s^{\mathcal{Q}}$, respectively $t^{\mathcal{Q}}$, will be used to denote that unique function $s$, respectively mapping $t$ such that the requirement from Lemma 8.7 is satisfied. The function $s^{\mathcal{Q}}$ will be called the *saturating function* (of the criterion $\mathcal{Q}$) and the mapping $t^{\mathcal{Q}}$ the $\ell$-standardized *transformation of data* (relative to $\mathcal{Q}$). Moreover, given a database $D \in \mathsf{DATA}(N, d)$, the vector $[t^{\mathcal{Q}}_D(A)]_{A \subseteq N} \in \mathbb{R}^{\mathcal{P}(N)}$ will be called the ($\ell$-standardized) *data vector* (relative to $\mathcal{Q}$).

CONSEQUENCE 8.5 Let $\mathcal{Q}$ be a regular quality criterion for learning DAG models and $K, L \in \mathsf{DAGS}(N)$ be such that $\mathcal{M}_K \sqsubset \mathcal{M}_L$. Moreover, let $t^{\mathcal{Q}}_D$ denote the data vector given by $D \in \mathsf{DATA}(N, d)$ and $u_{\langle a,b|C\rangle}$ be the differential imset for $K$ and $L$. Then

$$\mathcal{Q}(K, D) - \mathcal{Q}(L, D) \;=\; \langle t^{\mathcal{Q}}_D, u_{\langle a,b|C\rangle}\rangle\,. \tag{8.35}$$

**Proof:** This immediately follows from (8.31) and the definition of the differential imset $u_{\langle a,b|C\rangle} = u_L - u_K$. □

REMARK 8.13 The method of structural imsets leads to the following proposal of how to modify and implement the local search method for learning DAG models.

- The states of the search space can be represented by *standard imsets*.
- The moves between states of the space can be represented by *differential imsets*.
- Given a regular criterion $\mathcal{Q}$ for learning DAG models data can be represented by the respective $\ell$-standardized *data vector*.

The formula (8.31) says that $\mathcal{Q}$ is a linear function of those state representatives and data representatives. Moreover, owing to (8.35), the change in the value of $\mathcal{Q}$ after a move can be interpreted in terms of CI. Indeed, the value $\langle t_D^{\mathcal{Q}}, u_{\langle a,b|C\rangle}\rangle$ can be viewed as the qualitative evaluation of the step in which the hypothesis $a \perp\!\!\!\perp b \,|\, C$ is rejected. This interpretation is indeed possible in the case of the MLL criterion - see Remark 8.14. $\triangle$

OBSERVATION 8.4 The saturating function and the data vector relative to the MLL criterion are as follows:

$$\begin{array}{rcl} s^{\mathsf{MLL}}(D) & = & d\cdot H(\hat{P}|\upsilon)\,,\\ t_D^{\mathsf{MLL}}(A) & = & d\cdot m_{\hat{P}}(A)\,, \end{array} \qquad \text{for } D\in\mathsf{DATA}(N,d),\ A\subseteq N, \tag{8.36}$$

where $\hat{P}$ is the empirical measure computed from $D$ (see Section 10.10.1), $\upsilon$ the counting measure on $\mathsf{X}_N$ and $m_{\hat{P}}$ the multiinformation function (see Section 2.3.4).

**Proof:** Apply the procedure used in the proof of Lemma 8.7. By (8.32) and (8.29), and using Convention 3, write for $D\in\mathsf{DATA}(N,d)$

$$\boldsymbol{t}'_D(A) = \sum_{\boldsymbol{x}\in\mathsf{X}_A} d_{\boldsymbol{x}}\cdot\ln d_{\boldsymbol{x}} \qquad \text{if } \emptyset\neq A\subseteq N,\quad \boldsymbol{t}'_D(\emptyset) = d\cdot\ln d\,.$$

Hence, by (8.33)

$$\begin{array}{rcl} s^{\mathsf{MLL}}(D) & = & \displaystyle\sum_{\boldsymbol{x}\in\mathsf{X}_N} d_{\boldsymbol{x}}\cdot\ln d_{\boldsymbol{x}} - d\cdot\ln d = \sum_{\boldsymbol{x}\in\mathsf{X}_N} d_{\boldsymbol{x}}\cdot\ln d_{\boldsymbol{x}} - \sum_{\boldsymbol{x}\in\mathsf{X}_N} d_{xmx}\cdot\ln d = \\ & = & \displaystyle\sum_{\boldsymbol{x}\in\mathsf{X}_N} d_{\boldsymbol{x}}\cdot\ln\frac{d_{\boldsymbol{x}}}{d} = d\cdot\sum_{\boldsymbol{x}\in\mathsf{X}_N}\frac{d_{\boldsymbol{x}}}{d}\cdot\ln\frac{d_{\boldsymbol{x}}}{d} = d\cdot\sum_{\boldsymbol{x}\in\mathsf{X}_N}\hat{p}_N(\boldsymbol{x})\cdot\ln\hat{p}_N(\boldsymbol{x})\,. \end{array}$$

The following formula for $t_D^{\mathsf{MLL}}$ is implied by (8.34) using Convention 3:

$$\begin{array}{rcl} t_D^{\mathsf{MLL}}(A) & = & \displaystyle\sum_{\boldsymbol{x}\in\mathsf{X}_A} d_{\boldsymbol{x}}\cdot\ln d_{\boldsymbol{x}} - \sum_{i\in A}\sum_{\boldsymbol{y}\in\mathsf{X}_i} d_{\boldsymbol{y}}\cdot\ln d_{\boldsymbol{y}} + (|A|-1)\cdot d\cdot\ln d = \\ & = & \displaystyle\sum_{\boldsymbol{x}\in\mathsf{X}_A} d_{\boldsymbol{x}}\cdot\ln d_{\boldsymbol{x}} - \sum_{i\in A}\sum_{\boldsymbol{x}\in\mathsf{X}_A} d_{\boldsymbol{x}}\cdot\ln d_{\boldsymbol{x}_i} + (|A|-1)\cdot\sum_{\boldsymbol{x}\in\mathsf{X}_A} d_{\boldsymbol{x}}\cdot\ln d = \\ & = & \displaystyle\sum_{\boldsymbol{x}\in\mathsf{X}_A} d_{\boldsymbol{x}}\cdot\ln\frac{d_{\boldsymbol{x}}\cdot d^{|A|-1}}{\prod_{i\in A} d_{\boldsymbol{x}_i}} = d\cdot\sum_{\boldsymbol{x}\in\mathsf{X}_A}\frac{d_{\boldsymbol{x}}}{d}\cdot\ln\frac{d_{\boldsymbol{x}}\cdot d^{-1}}{\prod_{i\in A} d_{\boldsymbol{x}_i}\cdot d^{-1}} = \\ & = & \displaystyle d\cdot\sum_{\boldsymbol{x}\in\mathsf{X}_A}\hat{p}_A(\boldsymbol{x})\cdot\ln\frac{\hat{p}_A(\boldsymbol{x})}{\prod_{i\in A}\hat{p}_i(\boldsymbol{x}_i)} = d\cdot m_{\hat{P}}(A)\,. \end{array}$$

$\square$

Remark 8.14 It follows from Observation 8.4 and Consequence 2.2 that $t_D^{\mathsf{MLL}}$ is a supermodular function for every $D \in \mathsf{DATA}(N, d)$. More specifically, supposing $K, L \in \mathsf{DAGS}(N)$ such that $\mathcal{M}_K \sqsubset \mathcal{M}_L$ and $u_{\langle a,b|C\rangle}$ is the differential imset for $K$ and $L$, the formula (8.35) in Consequence 8.5 implies

$$\mathsf{MLL}\,(K, D) - \mathsf{MLL}\,(L, D) \;=\; \langle t_D^{\mathsf{MLL}}, u_{\langle a,b|C\rangle}\rangle = d \cdot \langle m_{\hat{P}}, u_{\langle a,b|C\rangle}\rangle\,.$$

Thus, the MLL criterion is descreasing with respect to the inclusion quasiordering. No matter what data are given, it attains its maximum for graphs whose underlying graph is complete (see Figure 8.1 for illustration). Moreover, it follows from the proof of Consequence 2.2 that the number $\langle m_{\hat{P}}, u_{\langle a,b|C\rangle}\rangle$ is just the relative entropy of $\hat{P}^{abC}$ with respect to the conditional product $\hat{Q}$ of $\hat{P}^{aC}$ and $\hat{P}^{bC}$ (see Remark 2.11 for this concept). Therefore,

$$\mathsf{MLL}\,(K, D) - \mathsf{MLL}\,(L, D) \;=\; d \cdot \langle m_{\hat{P}}, u_{\langle a,b|C\rangle}\rangle \;=\; d \cdot H(\hat{P}^{abC}|\hat{Q})\,.$$

Because $\hat{Q}$ is nothing but the fitted empirical measure defined in Section 10.10.1 the latter expression is one half of the value of the $G^2$-statistics for testing $a \perp\!\!\!\perp b\,|\,C$. This observation is interesting in connection with Whittaker's approach [148] to learning UG models by the minimization of the deviance (see Section 8.1). As a matter of fact, if one is interested in learning decomposable models which are known to be special cases of DAG models (see Section 3.4.1) then, by Remark 8.2, the value of the deviance of a statistical model described by an undirected graph $H$ over $N$ can be expressed as $\boldsymbol{k} - 2 \cdot \mathsf{MLL}\,(G, D)$ where $\boldsymbol{k}$ is a constant which does not depend on the model and $G \in \mathsf{DAGS}(N)$ induces the same model as $H$. Thus, the deviance difference for a pair of triangulated undirected graphs which differ in the presence of one line is nothing but the multiple by two of the difference in the value of the MLL criterion for a pair of acyclic directed graphs which are inclusion neighbours. In particular, as explained above, the deviance difference is the value of the $G^2$-statistic for testing the respective elementary CI statement. This explains the phenomenon observed by Whittaker [148] and mentioned on p. 137 and in Remark 8.2. △

Consequence 8.6 The saturating function and the data vector relative to the DIM criterion are as follows:

$$\begin{array}{rcl} s^{\mathsf{DIM}}(D) & = & \prod_{i\in N} r(i) - 1 \\ t_D^{\mathsf{DIM}}(A) & = & \prod_{i\in A} r(i) - \sum_{i\in A} r(i) + |A| - 1 \end{array} \qquad \text{for any } D \in \mathsf{DATA}(N, d),\ A \subseteq N\,. \tag{8.37}$$

The formulas for data vectors relative to the AIC criterion and the BIC criterion are here:

$$\begin{array}{rcl} t_D^{\mathsf{AIC}}(A) & = & d \cdot m_{\hat{P}}(A) - \prod_{i\in A} r(i) + \sum_{i\in A} r(i) - |A| + 1\,, \\ t_D^{\mathsf{BIC}}(A) & = & d \cdot m_{\hat{P}}(A) - \frac{\ln d}{2} \cdot \{\, \prod_{i\in A} r(i) - \sum_{i\in A} r(i) + |A| - 1 \,\}\,. \end{array} \tag{8.38}$$

**Proof:** The formula (8.37) can be obtained by substituting (8.30) into (8.33) and (8.34). The other formulas follow from Observation 8.4 using the definition of information criteria (see Section 10.10.4). □

Note that it follows from Consequence 8.6 that

$$\langle t_D^{\mathsf{DIM}}, u_{\langle a,b|C\rangle}\rangle \;=\; (r(a) - 1) \cdot (r(b) - 1) \cdot \prod_{c\in C} r(c) \qquad \text{for every } u_{\langle a,b|C\rangle} \in \mathcal{E}(N)\,.$$

In particular, $t_D^{\mathsf{DIM}}$ is also supermodular and the DIM criterion is decreasing with respect to inclusion quasiordering (see Figure 8.1 for illustration). On the other hand, this is not true for the AIC and BIC criteria: each of these two criteria is defined as a difference of two decreasing criteria. Thus, either of the criteria can attain its maximal values inside the Hasse diagram of the inclusion ordering for DAG models.

As indicated in Remark 8.9, I believe that formulas analogous to those in Consequence 8.6 can be derived for some of the Bayesian criteria for learning DAG models.

REMARK 8.15 One possible objection against the use of imsets is that memory demands for their representation in a computer are too high. However, if one is only interested in learning DAG models then standard imsets for acyclic directed graphs can be effectively represented in memory of a computer. Indeed, any standard imset over $N$ has at most $2|N|+2$ non-zero values - see (7.3). Thus, to represent it in computer memory, one can assign to every subset $A \subseteq N$ a numerical code and represent an imset $u$ by the list of, at most, $2 \cdot |N|+2$ items where each item is a pair $[code(A), u(A)]$ for a set $A \subseteq N$ with $u(A) \neq 0$. The number of bits needed to represent a standard imset is a polynomial function of $|N|$, which means that memory demands are equivalent to those in the case of the use of graphs over $N$.

As concerns computer representation of data, a data vector $t_D$ (see p. 159) can be represented by a list of $2^{|N|}$ items of the form $[code(A), t_D(A)]$ for $A \subseteq N$. This may appear to be even more effective than the traditional way of representing data in the form of a contingency table because $2^{|N|} \leq |X_N|$. On the other hand, a contingency table is an integral function on $X_N$ while $t_D$ is a real function on $\mathcal{P}(N)$. One can also keep a contingency table in the memory of a computer and compute any value of $t_D(A)$, $A \subseteq N$ each time this is needed. Thus, the memory demands for representing data are equal to those in the case of the use of graphs for representing models. $\triangle$

# Bibliography


[1] M. Aigner: Combinatorial Theory, Springer-Verlag 1979.

[2] H. Akaike: A new look at the statistical model identification, IEEE Transactions on Automatic Control 19 (1974), pp. 716-722.

[3] Z. An, D. A. Bell, J. G. Hughes: On the axiomatization of conditional independence, Kybernetes 21 (1992), n. 7, pp. 48-58.

[4] J. Anděl: Mathematical Statistics (in Czech), SNTL (Prague) 1985.

[5] S. A. Andersson, M. D. Perlman: Lattice models for conditional independence in multivariate normal distributions, Annals of Statistics 21 (1993), pp. 1318-1358.

[6] S. A. Andersson, D. Madigan, M. D. Perlman: A characterization of Markov equivalence classes for acyclic digraphs, Annals of Statistics 25 (1997), n. 2, pp. 505-541.

[7] S. A. Andersson, D. Madigan, M. D. Perlman, C. M. Triggs: A graphical characterization of lattice conditional independence models, Annals of Mathematics and Artificial Intelligence 21 (1997), pp. 27-50.

[8] S. A. Andersson, D. Madigan, M. D. Perlman: Alternative Markov properties for chain graphs, Scandinavian Journal of Statistics 28 (2001), n. 1, pp. 33-85.

[9] G. Birkhoff: Lattice Theory, AMS Colloquim Publications 1951.

[10] P. Boček: SGPOKUS, a computer program, Institute of Information Theory and Automation, June 1994.

[11] P. Boček: GENERATOR, a computer program, Institute of Information Theory and Automation, March 2001.

[12] R. R. Bouckaert: IDAGs: a perfect map for any distribution, in Symbolic and Quantitative Approaches to Reasoning and Uncertainty (M. Clarke, R. Kruse, S. Moral eds.), Lecture Notes in Computer Science 747, Springer-Verlag 1993, pp. 49-56.

[13] R. R. Bouckaert: Bayesian belief networks -from construction to inference, PhD thesis, University of Utrecht (Netherland) 1995.

[14] R. R. Bouckaert, M. Studený: Chain graphs: semantics and expressiveness, in Symbolic and Quantitative Approaches to Reasoning and Uncertainty (Ch. Froidevaux, J. Kohlas eds.), Lecture Notes in Artificial Intelligence 946, Springer-Verlag 1995, pp. 67-76.

[15] A. Brøndsted: An Introduction to Convex Polytopes, Springer-Verlag 1983 (Russian translation Mir 1988).

[16] L. M. de Campos, J. F. Huete, S. Moral: Probability intervals, a tool for uncertain reasoning, a technical report DECSAI-93206, July 1993, University of Granada.

[17] L. M. de Campos: Independency relationships in possibility theory and their application to learning belief networks, in Mathematical and Statistical Methods in Artificial Intelligence (G. Della Riccia, R. Kruse, R. Viertl eds.), Springer-Verlag 1995, pp. 119-130.

[18] L. M. de Campos: Characterization of decomposable dependency models, Journal of Artificial Intelligence Research 5 (1996), pp. 289-300.

[19] R. Castelo: The discrete acyclic digraph Markov model in data mining, PhD thesis, University of Utrecht 2002.

[20] J. Cheng, R. Greiner: Comparing Bayesian network classifiers, in Uncertainty in Artificial Intelligence 15 (K. B. Laskey, H. Prade eds.), Morgan Kaufmann 1999, pp. 101-107.

[21] D. M. Chickering: A transformational characterization of equivalent Bayesian network structures, in Uncertainty in Artificial Intelligence 11 (P. Besnard, S. Hanks eds.), Morgan Kaufmann, 1995, pp. 87-98.

[22] D. M. Chickering: Optimal structure identification with greedy search, Journal of Machine Learning Research 3 (2002), pp. 507-554.

[23] G. F. Cooper, E. Herskovits: A Bayesian method for induction of probabilistic networks from data, Machine Learning 9 (1992), pp. 309-341.

[24] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, D. J. Spiegelhalter: Probabilistic Networks and Expert Systems, Springer-Verlag 1999.

[25] D. R. Cox, D. V. Hinkley: Theoretical Statistics, Chapman and Hall 1982.

[26] D. R. Cox, N. Wermuth: Multivariate Dependencies - Models, Analysis and Interpretation, Chapman and Hall 1996.

[27] A. P. Dawid: Conditional independence in statistical theory, Journal of the Royal Statistical Society series B 41 (1979), n. 1, pp. 1-31.

[28] A. P. Dawid, S. L. Lauritzen: Hyper Markov laws in the statistical analysis of decomposable graphical models, Annals of Statistics 21 (1993), pp. 1272-1317.

[29] A. P. Dawid: Conditional independence, in Encyclopedia of Statistical Science Update vol. 2 (S. Kotz, C. B. Read, D. L. Banks eds.), John Wiley 1999, pp. 146-155.

[30] A. P. Dawid, M. Studený: Conditional products, an alternative approach to conditional independence, in Artificial Intelligence and Statistics 99, Proceedings of the 7th Workshop (D. Heckerman, J. Whittaker eds.), Morgan Kaufmann 1999, pp. 32-40.

[31] A. P. Dawid: Separoids: a mathematical framework for conditional independence and irrelevance, Annals of Mathematics and Artificial Intelligence 32 (2001), n. 1-4, pp. 335-372.

[32] A. P. Dempster: Covariance selection, Biometrocs 28 (1972), pp. 157-175.

[33] M. Fiedler: Special Matrices and Their Use in Numerical Mathematics (in Czech), SNTL (Prague) 1981.

[34] J.-P. Florens, M. Mouchart, J.-M. Rolin: Elements of Bayesian Statistics, Marcel Dekker 1990.

[35] N. Friedman, D. Geiger, M. Goldszmidt: Bayesian network classifier, Machine Learning 29 (1997), pp. 131-163.

[36] M. Frydenberg: The chain graph Markov property, Scandinavian Journal of Statistics 17 (1990), n. 4, pp. 333-353.

[37] M. Frydenberg: Marginalization and collapsibility in graphical interaction models, Annals of Statistics 18 (1990), n. 2, pp. 790-805.

[38] L. C. van der Gaag, J.-J. Ch. Meyer: Informational independence, models and normal forms, International Journal of Intelligent Systems 13 (1998), n. 1, pp. 83-109.

[39] B. Ganter, R. Wille: Formal Concepts Analysis - Mathematical Foundations, Springer-Verlag 1999.

[40] P. Giudici, P. J. Green: Decomposable graphical Gaussian determination, Biometrika 86 (1999), pp. 785-801.

[41] D. Geiger: Towards the formalization of informational dependencie, technical report CSD-850053, R-102, UCLA Los Angeles, December 1987.

[42] D. Geiger, T. Verma, J. Pearl: Identifying independence in Bayesian networks, Networks 20 (1990), n. 5, pp. 507-534.

[43] D. Geiger, J. Pearl: On the logic of causal models, in Uncertainty in Artificial Intelligence 4 (R. D. Shachter, T. S. Lewitt, L. N. Kanal and J. F. Lemmer eds.), North-Holland 1990, pp. 3-14.

[44] D. Geiger, A. Paz, J. Pearl: Axioms and algorithms for inferences involving probabilistic independence, Information and Computation 91 (1991), n. 1, pp. 128-141.

[45] D. Geiger, J. Pearl: Logical and algorithmic properties of conditional independence and graphical models, Annals of Statistics 21 (1993), n. 4, pp. 2001-2021.

[46] D. Geiger, A. Paz, J. Pearl: On testing whether an embedded Bayesian network represents a probability model, in Uncertainty in Artificial Intelligence 10 (R. L. de Mantaras, D. Poole eds.), Morgan Kaufmann 1994, pp. 244-252.

[47] P. Hájek, T. Havránek, R. Jiroušek: Uncertain Information Processing in Expert Systems, CRC Press 1992.

[48] D. Heckerman: A tutorial on learning Bayesian networks, technical report MSR-TR-95-06, Microsoft Research, Redmont, March 1995.

[49] R. A. Horn, Ch. R. Johnson: Matrix Analysis, Cambridge University Press 1986 (Russian translation, Mir 1989).

[50] R. Jiroušek: Solution of the marginal problem and decomposable distributions, Kybernetika 27 (1991), pp. 403-412.

[51] K. G. Jöreskog, D. Sörbom: LISREL 7 - A Guide to the Program and Application, SPSS Inc. 1989.

[52] G. Kauermann: On a dualization of graphical Gaussian models, Scandinavian Journal of Statistics 23 (1996), n. 1, pp. 105-116.

[53] H. G. Kellerer: Verteilungsfunktionen mit gegebenem Marginalverteilungen (in German), Z. Wahrscheinlichkeitstheorie 3 (1964), pp. 247-270.

[54] H. Kiiveri, T. P. Speed, J. B. Carlin: Recursive causal models, Journal of Australian Mathematical Society series A 36 (1984), pp. 30-52.

[55] T. Kočka: Graphical models - learning and applications. PhD thesis, University of Economics Prague (Czech Republic) 2001.

[56] T. Kočka, R. R. Bouckaert, M. Studený: On the inclusion problem, research report n. 2010, Institute of Information Theory and Automation, Prague, February 2001.

[57] J. T. A. Koster: Gibbs factorization and the Markov property, unpublished manuscript.

[58] J. T. A. Koster: Markov properties of nonrecursive causal models, Annals of Statistics 24 (1996), n. 5, pp. 2148-2177.

[59] J. T. A. Koster: Marginalizing and conditioning in graphical models, submitted to Bernoulli.

[60] I. Kramosil: A note on non-axiomatizability of independence relations generated by certain probabilistic structures, Kybernetika 24 (1988), n. 2, pp. 439-446.

[61] W. Lam, F. Bacchus: Learning Bayesian belief networks, an approach based on the MDL principle, Computational Intelligence 10 (1994), pp. 269-293.

[62] S. L. Lauritzen, N. Wermuth: Mixed interaction models, research report R-84-8, Inst. Elec. Sys., University of Aalborg 1984.
Note that this report was later modified and became a basis of the paper [65].

[63] S. L. Lauritzen, T. P. Speed, K. Vijayan: Decomposable graphs and hypergraphs, Journal of Australian Mathematical Society A 36 (1984), n. 1, pp. 12-29.

[64] S. L. Lauritzen, D. J. Spiegelhalter: Local computation with probabilities on graphical structures and their application to expert systems, Journal of the Royal Statistical Society series B 50 (1988), n. 2, pp. 157-224.

[65] S. L. Lauritzen, N. Wermuth: Graphical models for associations between variables, some of which are qualitative and some quantitative, Annals of Statistics 17 (1989), n. 1, pp. 31-57.

[66] S. L. Lauritzen: Mixed graphical association models, Scandinavian Journal of Statistics 16 (1989), n. 4, pp. 273-306.

[67] S. L. Lauritzen, A. P. Dawid, B. N. Larsen, H.-G. Leimer: Independence properties of directed Markov fields, Networks 20 (1990), n. 5, pp. 491-505.

[68] S. L. Lauritzen: Graphical Models, Clarendon Press 1996.

[69] E. L. Lehman: Testing Statistical Hypothesis, John Wiley 1957.

[70] H.-G. Leimer: Optimal decomposition by clique separators, Discrete Mathematics 113 (1993), pp. 99-123.

[71] M. Levitz, M. D. Perlman, D. Madigan: Separation and completeness properties for AMP chain graph Markov models, submitted to Annals of Statistics.

[72] M. Loéve: Probability Theory, Foundations, Random Processes, D. van Nostrand 1955.

[73] D. Madigan, J. York: Bayesian graphical models for discrete data, International Statistical Review 63 (1995), pp. 215-232.

[74] F. M. Malvestuto: A unique formal system for binary decomposition of database relations, probability distributions and graphs, Information Sciences 59 (1992), pp. 21-52. + F. M. Malvestuto, M. Studený: Comment on "A unique formal ... graphs", Information Sciences 63 (1992), pp. 1-2.

[75] F. M. Malvestuto: Theory of random observables in relational data bases, Information Systems 8 (1983), n. 4, pp. 281-289.

[76] J. L. Massey: Causal interpretation of random variables (in Russian), Problemy Peredachi Informatsii 32 (1996), n. 1, pp. 112-116.

[77] F. Matúš: Ascending and descending conditional independence relations, in Information Theory, Statistical Decision Functions and Random Processes, Transactions of 11th Prague Conference, vol. B (S. Kubík, J. Á. Víšek eds.), Kluwer 1992, pp. 189-200.

[78] F. Matúš: On equivalence of Markov properties over undirected graphs, Journal of Applied Probability 29 (1992), n. 3, pp. 745-749.

[79] F. Matúš: Probabilistic conditional independence structures and matroid theory, backgrounds, International Journal of General Systems 22 (1994), n. 2, pp. 185-196.

[80] F. Matúš: Stochastic independence, algebraic independence and abstract connectedness, Theoretical Computer Science A 134 (1994), n. 2, pp. 445-471.

[81] F. Matúš: On the maximum-entropy extensions of probability measures over undirected graphs, in Proceedings of WUPES'94, September 11-15, 1994, Třešť, Czech Republic, pp. 181-198.

[82] F. Matúš, M. Studený: Conditional independences among four random variables I., Combinatorics, Probability and Computing 4 (1995), n. 4, pp. 269-278.

[83] F. Matúš: Conditional independences among four random variables II., Combinatorics, Probability and Computing 4 (1995), n. 4, pp. 407-417.

[84] F. Matúš: Conditional independence structures examined via minors, Annals of Mathematics and Artificial Intelligence 21 (1997), pp. 99-128.

[85] F. Matúš: Conditional independences among four random variables III., final conclusion, Combinatorics, Probability and Computing 8 (1999), n. 3, pp. 269-276.

[86] F. Matúš: Lengths of semigraphoid inferences, Annals of Mathematics and Artificial Intelligence 35 (2002), pp. 287-294. ??????

[87] C. Meek: Causal inference and causal explanation with background knowlewdge, in Uncertainty in Artificial Intelligence 11 (P. Besnard, S. Hanks eds.), Morgan Kaufmann 1995, pp. 403-410.

[88] C. Meek: Strong completeness nad faithfulness in Bayesian networks, in Uncertainty in Artificial Intelligence 11 (P. Besnard, S. Hanks eds.), Morgan Kaufmann 1995, pp. 411-418.

[89] C. Meek: Graphical models, selecting causal and statistical models, PhD thesis, Carnegie Melon University 1997.

[90] E. Mendelson: Introduction to Mathematical Logic, 2nd edition, D. van Nostrand 1979.

[91] M. Mouchart, J.-M. Rolin: A note on conditional independence with statistical applications, Statistica 44 (1984), n. 4, pp. 557-584.

[92] M. Mouchart, J.-M. Rolin: Letter to the editor, Statistica 45 (1985), n. 3, pp. 427-430.

[93] J. Moussouris: Gibbs and Markov properties over undirected graphs, Journal of Statistical Physics 10 (1974), n. 1, pp. 11-31.

[94] J. Neveu: Bases Mathématiques du Calcul des Probabilités (in French), Masson et Cie 1964.

[95] A. Paz: A fast version of Lauritzen's algorithm for checking representation of independencies, in Proceedings of 5th Workshop on Uncertainty Processing (WUPES'00) June 21-24, 2000, J. Hradec, Czech Republic, p. 181-190.

[96] A. Paz, R. Y. Geva, M. Studený: Representation of irrelevance relations by annotated graphs, Fundamenta Informaticae 42 (2000), pp. 149-199.

[97] J. Pearl, A. Paz: Graphoids, graph-based logic for reasoning about relevance relations, in Advances in Artificial Intelligence II (B. Du Boulay, D. Hogg, L. Steels eds.), North-Holland 1987, pp. 357-363.

[98] J. Pearl: Probabilistic Reasoning in Intelligent Systems - Networks of Plausible Inference, Morgan Kaufmann 1988.

[99] A. Perez: $\varepsilon$-admissible simplifications of the dependence structure of a set of random variables, Kybernetika 13 (1979), pp. 439-449.

[100] M. D. Perlman, L. Wu: Lattice conditional independence models for contingency tables with non-monotone missing data pattern, Journal of Statistical Planning 79 (1999), pp. 259-287.

[101] C. van Putten, J. H. van Shuppen: Invariance properties of conditional independence relation, Annals of Probability 13 (1985), n. 3, pp. 934-945.
Note that by [92] the majority of the results of this paper is almost identical with the results of [91].

[102] C. R. Rao: Linear Statistical Inference and Its Application, John Wiley 1965.

[103] T. S. Richardson: A polynomial-time algorithm for deciding Markov equivalence of directed cyclic graphical models, in Uncertainty in Artificial Intelligence 12 (E. Horvitz, F. Jensen eds.), Morgan Kaufmann 1996, pp. 462-469.

[104] T. S. Richardson: A discovery algorithm for directed cyclic graphs, in Uncertainty in Artificial Intelligence 12 (E. Horvitz, F. Jensen eds.), Morgan Kaufmann 1996, pp. 454-461.

[105] T. S. Richardson, P. Spirtes: Ancestral graph Markov models, report n. 374, Department of Statistics, University of Washington Seattle (2001), submitted to Annals of Statistics.

[106] J. Rosenmüller, H. G. Weidner: Extreme convex set functions with finite carrier - general theory, Discrete Mathematics 10 (1974), n. 3-4, pp. 343-382.

[107] W. Rudin: Real and Complex Analysis, McGraw-Hill 1974.

[108] Y. Sagiv, S. F. Walecka: Subset dependencies and completeness result for a subclass of embedded multivalued dependencies, Journal of Association for Computing Machinery 29 (1982), n. 1, pp. 103-117.

[109] G. Shafer: Probabilistic Expert Systems, CBMS-NSF Regional Conference Series in Applies Mathematics 67, SIAM 1996.

[110] A. Schrijver: Theory of Linear and Integer Programming, John Wiley 1986.

[111] G. Schwartz: Estimating the dimension of a model, Annals of Statistics 6 (1978), pp. 461-464.

[112] L. S. Shapley: Cores of convex games, International Journal of Game Theory 1 (1971/1972), pp. 11-26.

[113] P. P. Shenoy: Conditional independence in valuation-based systems, International Journal of Approximate Reasoning 10 (1994), n. 3, pp. 203-234.

[114] P. P. Shenoy: Representing conditional independence relations by valuated networks, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 2 (1994), pp. 143-165.

[115] T. P. Speed: A note on nearest-neighbour Gibbs and Markov probabilities, Sankhyā series A 41 (1979), pp. 184-197.

[116] D. J. Spiegelhalter, S. L. Lauritzen: Sequential updating of conditional probabilities on directed graphical structures, Networks 20 (1990), n. 5, pp. 579-605.

[117] P. Spirtes, C. Glymour, R. Scheines: Causation, Prediction and Search, Lecture Notes in Statistics 81, Springer-Verlag 1993.

[118] P. Spirtes: Directed cyclic graphical representations of feedback models, in Uncertainty in Artificial Intelligence 11 (P. Besnard, S. Hanks eds.), Morgan Kaufmann 1995, pp. 491-498.

[119] W. Spohn: Stochastic independence, causal independence and shieldability, Journal of Philosophical Logic 9 (1980), n. 1, pp. 73-99.

[120] W. Spohn: On the properties of conditional independence, in Patrick Suppes, Scientific Philosopher vol. 1, Probability and Probabilistic Causality (P. Humphreys ed.), Kluwer 1994, pp. 173-196.

[121] J. Štěpán: Probability Theory - Mathematical Foundations (in Czech), Academia (Prague) 1987.

[122] M. Studený: Asymptotic behaviour of empirical multiinformation, Kybernetika 23 (1987), n. 2, pp. 124-135.

[123] M. Studený: Multiinformation and the problem of characterization of conditional independence relations, Problems of Control and Information Theory 18 (1989), n. 1, pp. 3-16.

[124] M. Studený: Convex set functions I. and II., research reports n. 1733 and n. 1734, Institute of Information Theory and Automation, Prague, November 1991.

[125] M. Studený: Conditional independence relations have no finite complete characterization, in Information Theory, Statistical Decision Functions and Random Processes, Transactions of 11th Prague Conference, vol. B (S. Kubík, J. Á. Víšek eds.), Kluwer 1992, pp. 377-396.

[126] M. Studený: Multiinformation and conditional independence II., research report n. 1751, Institute of Information Theory and Automation, Prague, September 1992.

[127] M. Studený: Formal properties of conditional independence in different calculi of AI, in Symbolic and Quantitative Approaches to Reasoning and Uncertainty (M. Clarke, R. Kruse, S. Moral eds.), Lecture Notes in Computer Science 747, Springer-Verlag 1993, pp. 341-348.

[128] M. Studený: Convex cones in finite-dimensional real vector spaces, Kybernetika 29 (1993), n. 2, pp. 180-200.

[129] M. Studený: Structural semigraphoids, International Journal of General Systems 22 (1994), n. 2, pp. 207-217.

[130] M. Studený, P. Boček: CI-models arising among 4 random variables, in Proceedings of WUPES'94, September 11-15, 1994, Třešť, Czech Republic, pp. 268-282.

[131] M. Studený: Description of structures of conditional stochastic independence by means of faces and imsets (a series of 3 papers), International Journal of General Systems 23 (1994/1995), n. 2-4, pp. 123-137, 201-219, 323-341.

[132] M. Studený: Semigraphoids and structures of probabilistic conditional independence, Annals of Mathematics and Artificial Intelligence 21 (1997), n. 1, pp. 71-98.

[133] M. Studený: A recovery algorithm for chain graphs, International Journal of Approximate Reasoning 17 (1997), n. 2-3, pp. 265-293.

[134] M. Studený: On marginalization, collapsibility and precollapsibility, in Distributions with Given Marginals and Moment Problems (V. Beneš, J. Štěpán eds.), Kluwer 1977, pp. 191-198.

[135] M. Studený, R. R. Bouckaert: On chain graph models for description of conditional independence structures, Annals of Statistics 26 (1998), n. 4, pp. 1434-1495.

[136] M. Studený: Bayesian networks from the point of view of chain graphs, in Uncertainty in Artificial Intelligence 14 (G. F. Cooper, S. Moral eds.), Morgan Kaufmann 1998, pp. 496-503.

[137] M. Studený: Complexity of structural models, in Prague Stochastics'98, August 23-28, Prague 1998, pp. 521-528.

[138] M. Studený, J. Vejnarová: The multiinformation function as a tool for measuring stochastic dependence, in Learning in Graphical Models (M. I. Jordan ed.), Kluwer 1998, pp. 261-298.

[139] M. Studený, R. R. Bouckaert, T. Kočka: Extreme supermodular set functions over five variables, research report n. 1977, Institute of Information and Automation, Prague, January 2000.

[140] M. Studený: On mathematical description of probabilistic conditional independence structures, DrSc thesis, Institute of Information and Automation, Prague, May 2001.

[141] J. Vejnarová: Conditional independence in possibility theory, in Proceedings of ISIPTA'99 (1st International Symposium on Imprecise Probabilities and their Applications) (G. de Cooman, F. G. Cozman, S. Moral, P. Walley eds.), pp. 343-351.

[142] I. Vajda: Theory of Statistical Inference and Information, Kluwer 1989.

[143] T. Verma, J. Pearl: Causal networks, semantics and expressiveness, in Uncertainty in Artificial Intelligence 4 (R. D. Shachter, T. S. Lewitt, L. N. Kanal and J. F. Lemmer eds.), North-Holland 1990, pp. 69-76.

[144] T. Verma, J. Pearl: Equivalence and synthesis of causal models, in Uncertainty in Artificial Intelligence 6 (P. P. Bonissone, M. Henrion, L. N. Kanal, J. F. Lemmer eds.), Elsevier 1991, pp. 220-227.

[145] T. Verma, J. Pearl: An algorithm for deciding if a set of observed independencies has a causal explanation, in Uncertainty in Artificial Intelligence 8 (D. Dubois, M. P. Wellman, B. D'Ambrosio, P. Smets eds.), Morgan Kaufmann 1992, pp. 323-330.

[146] M. Volf, M. Studený: A graphical characterization of the largest chain graphs, International Journal of Approximate Reasoning 20 (1999), n. 3, pp. 209-236.

[147] S. Watanabe: Information theoretical analysis of multivariate correlation, IBM Journal of Research and Development 4 (1960), pp. 66-81.

[148] J. Whittaker: Graphical Models in Applied Multivariate Statistics, John Wiley 1990.

[149] Y. Xiang, S. K. M. Wong, N. Cercone: Critical remarks on single link search in learning belief networks, in Uncertainty in Artificial Intelligence 12 (E. Horvitz, F. Jensen eds.), Morgan Kaufmann 1996, pp. 564-571.