Mathematical aspects of learning Bayesian networks: Bayesian quality criteria

Milan Studený*
Institute of Information Theory and Automation of the ASCR
email: studeny@utia.cas.cz

December 29, 2008

Abstract

The motivation for this research report is learning a Bayesian network (BN) structure by the method of maximizing a quality criterion. The aim is to summarize the mathematical grounding for the Bayesian approach to learning a BN structure. At first, some of basic statistical concepts are recapitulated. Then the classes of *multinomial* and *Dirichlet* distributions are dealt with in more detail. A peculiar question what is, in fact, the correct dominating measure for (the class of) Dirichlet distributions is answered. After that basic Bayesian terminology is recalled and the (statistical) model of a discrete BN is formally introduced. It is shown to be an exponential family. This allows one to introduce a Bayesian model for (learning discrete) BN structures, including explicit specification of the mathematical assumptions taken from the literature. This leads to the formula for the (data vector of the) corresponding Bayesian quality criterion (= the logarithm of the marginal likelihood).

Keywords: learning Bayesian network structure; statistical model; multinomial distribution; Dirichlet distribution; exponential family; Bayesian quality criterion; data vector.

1 Introduction

The general motivation for this technical report is learning a *Bayesian network* (BN) structure by the method of maximizing a *quality criterion*, often named *score metric*, *scoring criterion* or simply *score* by other authors [6, 22, 7]. More specifically, what is meant is an algebraic approach to learning Bayesian networks, in which every BN structure is represented by a special integral vector, called the *standard imset* and the data(base) is encoded in the form of a special real vector (of the same dimension), called the *data vector*. This approach is described in more detail in Chapter 8 of [27].

A popular class of quality criteria (for learning a BN structure) is the class of *Bayesian criteria*, also referred as the (*logarithm of the*) *marginal likelihood* in the literature [9]. A correct mathematical definition of a quality criterion of this type is based on a series of

*This research has been supported by the grant GAČR n. 201/08/0539.

special technical assumptions. In the literature on this topic (the author of this report has an opportunity to consult) these assumptions are either not mentioned explicitly or formulated in a vague way, not in clear mathematical terms.

Moreover, some natural mathematical questions have been omitted in those papers and books, perhaps intentionally. The reasons could be different: experts in Bayesian statistics may consider them to be evident from an intuitive point of view, while computer scientists may simply wish to skip technical details that are not important from their point of view. However, these questions are quite important from the point of view of a mathematician. One has to clarify these things in order to be able to deal with deeper subsequent mathematical questions.

This technical report is a kind of a review directed towards special assumptions of the Bayesian approach (to learning a BN structure). It is an attempt to re-formulate these assumptions consistently in mathematical terms. Therefore, the report involves (some of) the definitions of elementary statistical concepts, which are well-known in statistical community (and, typically, omitted in common statistical papers).

The goal is to summarize basic concepts and assumptions of the Bayesian approach to learning a (discrete) BN structure. These assumptions have to be clarified in order to establish the basis for reliable research in deeper mathematical questions related to this approach. For example, the future research can deal with:

- the asymptotic behavior of Bayesian quality criteria, which is related to the question of their statistical consistency,
- the geometric aspects of the task to maximize a quality criterion of this kind.

2 Preliminaries

This is to review a few elementary statistical concepts, just for the purpose of this report. The discrete case is only considered.

2.1 Sample space

In probability theory, by the *sample space* is usually meant the set of all possible outcomes (= results) of an experiment or of a series of experiments. Thus, from the point of view of statistics, the sample space is the set of possible values for the data. However, from the mathematical point of view, there are three different levels of generality/understanding for this concept.

A. Single outcome In this case, the elements of the sample space are possible outcomes of one single experiment/measurement. Thus, the sample space coincides with the set of all outcomes, called the *outcome space* (for a single experiment) in this report. Since the discrete case is only considered here, the outcome space X will be a non-empty

finite set throughout this report.[1] In the case of the discrete BN model described later in § 6, the outcome space X has a special form (= internal structure): it is the Cartesian product $\prod_{i\in N} \mathsf{X}_i$ of certain (non-empty) finite sets X_i assigned to (= indexed by) elements of the set N of variables in consideration.

The most appropriate name for the probability distribution on the outcome space X, which is meant to have the crucial role in a "data-generating process", is probably the *theoretical distribution*.

B. Sample = database In this case, the elements of the sample space are ordered (finite) sequences of possible outcomes of successive experiments/measurements. This corresponds to a series of experiments. If the sequence has the length $d \geq 1$ then an element of this space is called a *sample of the size* d. Note that some computer scientists prefer the phrase a *database of the length* d instead.

Mathematically, the sample space is now the d-th power $X^{\{1,\dots,d\}}$ of the outcome space, that is, the collection of all mappings from $\{1,\dots,d\}$ to the outcome space X.

In this report, the probability distribution on this sample space will be named *sampling distribution*, in order to distinguish it from the (theoretical) distribution on the outcome space X.[2]

C. Table of counts = contingency table In this case, the elements of the sample space are possible tables of counts (of particular outcomes in a sample). Of course, this approach has reasonable sense in the discrete case only. Formally, the *table of counts* corresponding to a sample $y^1,\dots,y^d$ of the size $d \geq 1$ is a function c from the outcome space X to $\{0,1,\dots,d\}$, which ascribes to every $x \in X$ the number $c(x)$ of its occurrences in the sample: $c(x) = |\{j;\, y^j = x\}|$ for $x \in X$. Some people also use the phrase *contingency table* instead in the situation the outcome space X has the special form $\prod_{i\in N} \mathsf{X}_i$ with a finite set N of variables. Mathematically, the sample space is now the collection of all mappings $c : X \to \{0,1,\dots,d\}$ such that $\sum_{x\in X} c(x) = d$.

The corresponding distribution on this space is then the *multinomial distribution*, discussed in more detail in § 3.2.

In this report, the standard symbol for the *sample space* will be $\mathbb{X}$. Depending on the considered situation (= one of three cases mentioned above) one can have:

- $\mathbb{X} = X$, where X is the outcome space, or
- $\mathbb{X} = X^{\{1,\dots,d\}}$ for some $d \in \mathbb{N}$, or
- $\mathbb{X} = \{\, c \in \{0,1,\dots,d\}^X;\ \sum_{x\in X} c(x) = d \,\}$.

[1] It is natural to assume that X has at least two elements, that is, at least two different outcomes of the experiment are possible.

[2] Note that the phrase *sampling distribution* has often wider meaning. It is also used to name the distribution of a *statistic*, that is, of a (measurable) function on $X^{\{1,\dots,d\}}$.

Since the (basic) outcome space X will be assumed to be finite throughout this report, the corresponding sample space $\mathbb{X}$ will be finite as well, in all these three cases.

In general, however, statisticians may also consider an infinite sample space, typically (a special subset of) the real Euclidean space $\mathbb{R}^n$, $n \geq 1$. Then the sample space $\mathbb{X}$ is, moreover, endowed with a σ-algebra $\mathcal{X}$ of its subsets, typically the σ-algebra of Borel subsets (in the corresponding Euclidean/metric space). Thus, the sample space becomes a *measurable space* $(\mathbb{X}, \mathcal{X})$, and this mental step allows one to use tools of measure theory.

Of course, in the case of a finite sample space $\mathbb{X}$ mentioned above, the σ-algebra $\mathcal{X}$ will always be the collection of all subsets of $\mathbb{X}$.[3]

2.2 Statistical model and exponential family

In statistics, the situation when the probability distribution on the sample space is only partially known is in modelled by means of the concept of a statistical model.

Definition 1 (statistical model)
Let $\mathbb{X}$ be a sample space, endowed with a σ-algebra of subsets $\mathcal{X}$. By a *statistical model* will be meant a parameterized class of distributions on the measurable space $(\mathbb{X}, \mathcal{X})$:

$$\mathcal{P} = \{P_\theta;\, \theta \in \Theta\}\,.$$

The set Θ will be called the *parameter space* then.

What is assumed quite often is that the parameter space Θ is endowed with a σ-algebra $\mathcal{A}$ of its subsets and the class $\mathcal{P}$ is a *Markov kernel* from $(\Theta, \mathcal{A})$ to $(\mathbb{X}, \mathcal{X})$, that is,

$$\forall\, S \in \mathcal{X} \qquad \theta \mapsto P_\theta(S) \text{ is } \mathcal{A}\text{-measurable mapping.}$$

The parameter space Θ is typically an open connected subset in an (affine) Euclidean space and $\mathcal{A}$ is the class of Borel sets in it then.

More specifically, very common assumption on the statistical model is that it is an exponential family. The following definition can be found either in § 2.7 of [20] or in § XIV.3 of [1].

Definition 2 (exponential family)
Let $(\Theta, \mathcal{A})$ be a parameter space and $(\mathbb{X}, \mathcal{X})$ a sample space. A class $\mathcal{P} = \{P_\theta;\, \theta \in \Theta\}$ will be called an *exponential family* (of distributions) if there exists a σ-finite measure μ on $(\mathbb{X}, \mathcal{X})$ such that

$$\forall\, \theta \in \Theta \qquad P_\theta \ll \mu\,,$$

[3] Later, in the connection with the Bayesian approach, we will also consider infinite sets in place of sample spaces, but we will not denote them by $\mathbb{X}$ because of their different role in the joint global model - see § 4 and § 5 for details.

that is, μ is a *dominating measure* for $\mathcal{P}$, and, moreover, the densities of distributions can be expressed in the form

$$\frac{dP_\theta}{d\mu}(x) \equiv p_\theta(x) = c(\theta) \cdot u(x) \cdot \exp\,(\sum_{s=1}^{m} q_s(\theta) \cdot t_s(x)\,) \quad \text{for} \quad x \in \mathbb{X}\,,\ \theta \in \Theta\,, \tag{1}$$

where one has

- $m \in \mathbb{N},\ \ q_1, \ldots, q_m : \Theta \to \mathbb{R},\ \ t_1, \ldots, t_m : \mathbb{X} \to \mathbb{R},$
- $u : \mathbb{X} \to [0, +\infty),\ \ c : \Theta \to (0, +\infty),$

are all (correspondingly) measurable functions.

The value $c(\theta)$ for $\theta \in \Theta$ is the *normalizing constant* (for p_θ). The vector function $t(x) = [t_1(x), \ldots t_m(x)]$ is then a *sufficient statistic* for the class $\mathcal{P} = \{P_\theta;\, \theta \in \Theta\}$. This means the conditional distribution on $\mathbb{X}$ given the value of t does not depend on θ.[4] As mentioned in § 1.9 of [20] (see also § XV.5 of [1]), to verify that one can use the factorization criterion of sufficiency, which requires $p_\theta(x)$ can be written as follows:

$$p_\theta(x) = g(\theta, t(x)) \cdot h(x) \qquad \text{for} \quad \theta \in \Theta,\, x \in \mathbb{X}\,,$$

where g and h are (correspondingly) measurable.
Indeed, we put $g(\theta, y) = c(\theta) \cdot \exp\,(\sum_{s=1}^{m} q_s(\theta) \cdot y_s)$ for $y \equiv [y_1, \ldots, y_m]$ and $h(x) \equiv u(x)$.

Remark Note that the terminology concerning exponential families is not completely unified. Some people understand by an exponential family a class of distributions of the form (1), where Θ is a subset of $\mathbb{R}^m$ and $q_1, \ldots, q_m$ are coordinate functions.[5] To name the situation the actual dimension of the set of parameters Θ is smaller than the number m of components of the vector statistic they use the phrase a "*curved exponential family*".

3 Discrete statistical models

Basic discrete *statistical model for a single outcome*, that is, a model for discrete (strictly) positive *theoretical distributions*, can be introduced as follows:

Fixed parameter: $r \geq 1$.

The meaning of this constant parameter is the number of possible outcomes.

Sample space: $X = \{a_1, \ldots a_r\}$.

This is a finite set of outcomes/results of a certain experiment/measurement (= the outcome space). The k-th outcome is denoted by a_k here.

[4] The intuitive meaning is that (all) the essential information about the value of the parameter $\theta \in \Theta$ that can be read from the value x in the sample space $\mathbb{X}$ is brought by the value $t(x)$ of the sufficient statistic.

[5] Some of them then implicitly assume that Θ is convex or modify (= extend) Θ to get a convex set.

Parameter space: $\Theta = \{\, \theta \equiv (\theta_1, \ldots, \theta_r)\,;\ \theta_k > 0,\ \sum_{k=1}^{r} \theta_k = 1 \,\}$.

Having r fixed, the parameter space is the interior of so-called *probability simplex* in $\mathbb{R}^r$, that is, of $\{\theta;\, \theta_k \geq 0,\ \sum_{k=1}^{r} \theta_k = 1\,\}$. Because of the functional dependence $\sum_{k=1}^{r} \theta_k = 1$ between vector components (= single parameters) the actual number of free parameters in Θ is $r-1$; it is an open set in the corresponding affine space.

The formula for the density of P_θ (with respect to the arithmetic measure on X):

$$\forall\, \theta \in \Theta \qquad p_\theta(a_k) = \theta_k \ \text{ for } k = 1, \ldots, r\,.$$

The value of the density for the k-th outcome a_k is the k-th component of the vector parameter θ. Thus, in fact, the single parameters themselves are the theoretical probabilities of particular outcomes.

It is straightforward that this defines an exponential family, c.f. (1):

- μ is the arithmetic measure on X, $m = r$, $u \equiv 1$, $c \equiv 1$,
- $q_s(\theta) = \ln \theta_s$ for $s = 1, \ldots, m$,
- $t_s(x) = 1$ if $x = a_s$, $t_s(x) = 0$ otherwise.[6]

3.1 Sampling distribution

One of the central concepts in statistics is that of a random sample. Specifically, by a *random sample* of the size $d \geq 1$ from a (theoretical) distribution P is meant the sequence $\xi_1, \ldots, \xi_d$ of random variables that are independent and identically distributed, with shared distribution P. The distribution P is usually unknown, but it is assumed to belong to a statistical model (for theoretical distributions). Then, our knowledge about the (joint) distribution of the random sample can also be described in this way.

In other words, the previously mentioned statistical model for a single (discrete) outcome induces a *statistical model for a sample* of the size $d \geq 1$. The parameter space is the same, but the sample space is already the d-th power of the outcome space. The corresponding *sampling distribution* is the d-multiple product of the theoretical distribution. More specifically, the statistical model is given as follows:

Fixed parameters: $d, r \in \mathbb{Z}$, $d, r \geq 1$.

The meaning of the parameter d is the size of the sample, r is again the number of (possible) outcomes.

[6] This simple parameterization, which is a basis for later parameterization of the statistical model of a BN structure in § 6, is "symmetric" relative to the outcomes (= elements of X). On the other hand, it has the property that the actual dimension of Θ is less than m. One can provide an alternative parameterization of the same statistical model for positive theoretical distributions in which $m = r - 1$ coincides with the dimension of the corresponding parameter space. For example, $\theta_s \sim \ln \frac{p(a_s)}{p(a_r)}$ for $s = 1, \ldots, r-1$, but this alternative parameterization is not "symmetric" because it has a distinguished outcome a_r.

Sample space: $\mathbb{X} = X^{\{1,\dots,d\}}$, where $X = \{a_1, \dots a_r\}$.

The set X is a finite set of outcomes/results of a certain experiment/measurement (= the outcome space). The k-th outcome is denoted by a_k. The sample space $\mathbb{X}$ is the collection of samples of the size d, that is, of (ordered finite) sequences of outcomes of the length d.

Parameter space: $\Theta = \{\, \theta \equiv (\theta_1, \dots, \theta_r)\,;\ \theta_k > 0,\ \sum_{k=1}^{r} \theta_k = 1 \,\}$.

The parameter space is again the interior of the probability simplex in $\mathbb{R}^r$.

The formula for the density (with respect to the arithmetic measure on $\mathbb{X}$):

$$\forall\, \theta \in \Theta \ \ \forall\, y \equiv [y^1, \dots, y^d] \in \mathbb{X} \qquad p_\theta(y) = \theta_1^{x_1} \cdot \ldots \cdot \theta_r^{x_r}\,, \tag{2}$$

where x_k denotes the number of occurrences of a_k in y (for $k = 1, \dots, r$).

For each $\theta \in \Theta$, the sampling distribution on $\mathbb{X}$ is nothing but the d-multiple product of the corresponding theoretical distribution on X.

The formula (2) for the density can be derived as follows:
The theoretical probability of a_k is $p^*_\theta(a_k) = \theta_k$ for $k = 1, \dots, r$. If $\xi_1(\omega), \dots, \xi_d(\omega)$ is a random sample from this theoretical distribution, then the probability of occurrence of $y \equiv [y^1, \dots, y^d]$ is

$$Prob\,(\{\omega;\, [\xi_1(\omega), \dots, \xi_d(\omega)] = [y^1, \dots, y^d]\,\}) = \prod_{\ell=1}^{d} Prob\,(\{\omega;\, \xi_\ell(\omega) = y^\ell\,\}) = \prod_{\ell=1}^{d} p^*_\theta(y^\ell)\,.$$

Now, by the definition of x_1, the term $p^*_\theta(a_1) \equiv \theta_1$ occurs x_1-times in the last product. This gives a contribution $\theta_1^{x_1}$. Analogously, θ_2 occurs x_2-times, etc. Hence, the last product is nothing but $\theta_1^{x_1} \cdot \ldots \cdot \theta_r^{x_r}$. Therefore, this is the value of the sampling density for $y \equiv [y^1, \dots, y^d]$.

Again, it is easy to see that, having $d, r \geq 1$ fixed, the above class of sampling distributions defines an exponential family:

- μ is the arithmetic measure on $\mathbb{X} \equiv X^{\{1,\dots,d\}}$ with $X = \{a_1, \dots a_r\}$,
- $m = r$,
- $c(\theta) \equiv 1$ for $\theta \in \Theta$, $\quad u(y) \equiv 1$ for $y \in \mathbb{X}$,
- $q_s(\theta) = \ln \theta_s$ for $\theta \in \Theta$ and $s = 1, \dots, m$,
- $t_s(y) = x_s \equiv |\{\ell;\, 1 \leq \ell \leq d,\ y^\ell = a_s\}|$ for $y = [y^1, \dots, y^d] \in \mathbb{X}$.

More specifically, let us substitute to the formula (1):

$$\begin{aligned} p_\theta(y) &= c(\theta) \cdot u(y) \cdot \exp\,(\sum_{s=1}^{m} t_s(y) \cdot q_s(\theta)\,) = 1 \cdot 1 \cdot \exp\,(\sum_{s=1}^{r} x_s \cdot \ln \theta_s\,) \\ &= \exp\,(\sum_{s=1}^{r} \ln \theta_s^{x_s}\,) = \prod_{s=1}^{r} \exp\,(\ln\, \theta_s^{x_s}\,) = \prod_{s=1}^{r} \theta_s^{x_s} = \theta_1^{x_1} \cdot \ldots \cdot \theta_r^{x_r}\,, \end{aligned}$$

which is the expression in (2).

In particular the vector function $t : \mathbb{X} \equiv \{a_1, \dots a_r\}^{\{1,\dots,d\}} \to \mathbb{R}^r$ given by:

$$t_s(y) \;=\; x_s \;\equiv\; |\{\ell;\, 1 \le \ell \le d,\; y^\ell = a_s\}| \qquad \text{for } s = 1, \dots, r \;\text{ and }\; y \equiv [y^1, \dots, y^d] \in \{a_1, \dots a_r\}^{\{1,\dots,d\}}\,, \tag{3}$$

defines a *sufficient statistic* for this statistical model. Observe that $t(y) = [t_1(y), \dots, t_r(y)]$ is nothing but the table of counts corresponding to the sample $y = [y^1, \dots, y^d]$.

Terminological remark Some authors in computer science [16, 17, 22] use a special phrase "*multinomial sample*" to name (a sequence of random variables with the) distribution on $\mathbb{X} \equiv X^{\{1,\dots,d\}}$ given by (2). They were probably inspired by (a single sentence from a book by) Good [14], who used this phrase when he tried to relate this sampling distribution to the classic multinomial distribution for the corresponding tables of counts (see § 3.2).

In my view, using the adjective "*multinomial*" is inappropriate and misleading in connection with distributions on $\mathbb{X} \equiv X^{\{1,\dots,d\}}$. Of course, this adjective perfectly fits the distributions on (possible) tables of counts since the corresponding formula (4) below contains the multinomial coefficient; but there is no good reason for using this adjective in the context of the distributions for samples. That's why I use the phrase "sampling distribution" in this report instead.

Actually, I have found out that some young researchers in computer science, probably inspired by the above mentioned authors, have tried to use the phrase "*multinomial distribution*" to name a (strictly positive) theoretical distribution on $X = \{a_1, \dots a_r\}$. Again, there is no reason for the use of the adjective "multinomial" here, and, even worse, this is in direct contrast with common statistical terminology! The reader can learn in any basic (text)book on statistics that by a multinomial distribution is meant a special discrete distribution for tables of counts – see below.

3.2 Multinomial distribution

The multinomial distribution is one of the most important discrete multidimensional distributions in statistics. It can be interpreted as the distribution on (the collection of all possible) tables of counts corresponding to a random sample (from a strictly positive discrete theoretical distribution).

In other words, the statistical model for a sample (of the size $d \ge 1$) mentioned in § 3.1 induces a *statistical model for tables of counts*. The parameter space is the same, but the sample space is now the collection of all possible tables of counts. More specifically, the class of *multinomial distributions* can be introduced as follows (c.f. § XI.1. [1]):

Fixed parameters: $d, r \in \mathbb{Z}$, $d, r \ge 1$.
The meaning of the parameter d is the number of trials (= sample size), r is the number of (possible) outcomes.

Sample space: $\mathbb{X} = \{\, [x_1, \dots, x_r]\,;\; x_k \in \{0, \dots, d\} \;\; \sum_{k=1}^r x_k = d \,\} \subseteq \mathbb{R}^r$.
Thus, one deals with r-dimensional discrete distributions. The sample space can be interpreted as the collection of tables of counts for the outcome space $X = \{1, \dots, r\}$ and the sample size d. Of course, it is only a formal difference if one has $X = \{a_1, \dots, a_r\}$ instead (as in the previous sections).

Parameter space: $\Theta = \{\, \theta \equiv (\theta_1, \dots, \theta_r)\,;\; \theta_k > 0,\; \sum_{k=1}^r \theta_r = 1 \,\}$.
The parameter space is again the interior of the probability simplex in $\mathbb{R}^r$.

The formula for the density (with respect to the arithmetic measure on $\mathbb{X}$):

$$\forall\, \theta \equiv (\theta_1, \ldots, \theta_r) \in \Theta \quad \forall\, x \equiv [x_1, \ldots, x_r] \in \mathbb{X}$$
$$p_\theta(x) = \frac{d!}{x_1! \cdot \ldots \cdot x_r!} \cdot \theta_1^{x_1} \cdot \ldots \cdot \theta_r^{x_r}\,, \tag{4}$$

The formula follows from the above mentioned interpretation of multinomial distribution. The name of the distribution is clearly motivated by the *multinomial coefficient* $\binom{d}{x_1 \ldots x_r} \equiv \frac{d!}{x_1! \cdot \ldots \cdot x_r!}$ which occurs in the formula (4).

More specifically, the formula (4) can be derived as follows:
Let $X = \{a_1, \ldots, a_r\}$ be the outcome space. If $\theta \equiv (\theta_1, \ldots, \theta_r) \in \Theta$ defines the theoretical distribution $p^*_\theta(a_k) = \theta_k$ for $k = 1, \ldots, r$, then the formula (2) gives the probability of occurrence of a sample (= database) $y \equiv [y^1, \ldots, y^d] \in X^{\{1,\ldots,d\}}$. It suffices to find out which of these samples correspond to the table of counts $x \equiv [x_1, \ldots, x_r] \in \mathbb{X}$ and sum their occurrence probabilities. Nevertheless, by (2), all these samples have the same probability $\theta_1^{x_1} \cdot \ldots \cdot \theta_r^{x_r}$. Thus, to get the formula (4) is suffices to verify that the number of these samples (= databases) is $\frac{d!}{x_1! \cdot \ldots \cdot x_r!}$.
Indeed, for the choice of the positions of the occurrence of a_1 in $[y^1, \ldots, y^d]$ one has $\frac{d!}{x_1! \cdot (d-x_1)!}$ options, because this is equivalent to the choice of an x_1-element subset of $\{1, \ldots, d\}$. Then an $(d - x_1)$-element subset remains and for the choice (of the positions) of a_2 one has as many options as the number of x_2-element subsets of this $(d - x_1)$-element set, that is, $\frac{(d-x_1)!}{x_2! \cdot (d-x_1-x_2)!}$ options. By multiplying we get the overall number of options for positions of both a_1 and a_2:

$$\frac{d!}{x_1! \cdot (d-x_1)!} \cdot \frac{(d-x_1)!}{x_2! \cdot (d-x_1-x_2)!} = \frac{d!}{x_1! \cdot x_2! \cdot (d-x_1-x_2)!}\,.$$

We proceed by induction and, in the end, get the desired value $\frac{d!}{x_1! \cdot \ldots \cdot x_r!}$.

Again, it is easy to observe that, having $d, r \geq 1$ fixed, the class of multinomial distributions is an exponential family:

- μ is the arithmetic measure on $\mathbb{X} \equiv \{\, [x_1, \ldots, x_r]\,;\ x_k \in \{0, \ldots, d\}\ \sum_{k=1}^r x_k = d\}$,
- $m = r$,
- $c(\theta) \equiv 1$ for $\theta \in \Theta$,
- $u(x) = \dfrac{d!}{x_1! \cdot \ldots \cdot x_r!}$ for $x \equiv [x_1, \ldots, x_r] \in \mathbb{X}$,
- $q_s(\theta) = \ln \theta_s$ for $\theta \in \Theta$, $s = 1, \ldots, r$,
- $t_s(x) = x_s$ for $x \in \mathbb{X}$, $s = 1, \ldots, r$.

To check that let us substitute to the formula (1):

$$\begin{aligned} p_\theta(x) &= c(\theta) \cdot u(x) \cdot \exp\big(\sum_{s=1}^m t_s(x) \cdot q_s(\theta)\big) = 1 \cdot \frac{d!}{x_1! \cdot \ldots \cdot x_r!} \cdot \exp\big(\sum_{s=1}^r x_s \cdot \ln \theta_s\big) \\ &= \frac{d!}{x_1! \cdot \ldots \cdot x_r!} \cdot \exp\big(\sum_{s=1}^r \ln \theta_s^{x_s}\big) = \frac{d!}{x_1! \cdot \ldots \cdot x_r!} \cdot \prod_{s=1}^r \exp\big(\ln \theta_s^{x_s}\big) \\ &= \frac{d!}{x_1! \cdot \ldots \cdot x_r!} \cdot \prod_{s=1}^r \theta_s^{x_s} = \frac{d!}{x_1! \cdot \ldots \cdot x_r!} \cdot \theta_1^{x_1} \cdot \ldots \cdot \theta_r^{x_r}\,, \end{aligned}$$

which is the expression in (4).

Well-known formulas for the expectation vector and the covariance matrix of a random vector $[\zeta_1, \ldots, \zeta_r]$ with multinomial distribution are as follows (see § XI.1, Theorem 2 in [1]):

$$E(\zeta_k) = d \cdot \theta_k, \qquad var(\zeta_k) = d \cdot \theta_k \cdot (1 - \theta_k) \quad \text{for } k \in \{1, \ldots r\},$$
$$cov(\zeta_k, \zeta_l) = -d \cdot \theta_k \cdot \theta_l \quad \text{for } k \neq l,\ k, l \in \{1, \ldots r\}.$$

4 Dirichlet distribution

Dirichlet distribution is a very important continuous multidimensional distribution. It is a kind of standard distribution on the parameter space Θ for the discrete statistical models from the previous section. The class of Dirichlet distributions on Θ can be viewed as a statistical model, too.

4.1 Sample space and the correct dominating measure on it

Thus, the sample space for (the class of) Dirichlet distributions is nothing but the interior of the probability simplex in $\mathbb{R}^r$, $r \geq 2$:

$$\Theta = \{\, \theta \equiv (\theta_1, \ldots, \theta_r)\,;\ \theta_k > 0,\ \sum_{k=1}^{r} \theta_k = 1 \,\},$$

which is the parameter space in all three cases mentioned in § 3.

However, before giving the formula(s) for the densities of Dirichlet distributions one should specify with respect to which dominating measure the densities are considered. The affine hull of Θ is the set

$$\mathsf{aff}(\Theta) = \{\, \theta \equiv (\theta_1, \ldots, \theta_r)\,;\ \sum_{k=1}^{r} \theta_k = 1 \,\}.$$

It is an affine (= shifted linear) subspace of $\mathbb{R}^r$, of the dimension $r - 1$. Therefore, one can define consistently the (concept of the proper) Lebesgue measure on $\mathsf{aff}(\Theta)$ (see § B.3, Definition 14 and Proposition 14, for details). However, the usual standard formula for the density of the Dirichlet distribution given below is not meant with respect to the restriction of this proper Lebesgue measure on $\mathsf{aff}(\Theta)$ to Θ, but with respect to its $\frac{1}{\sqrt{r}}$-multiple!

That measure is the restriction (to Θ) of the image of the standard $(r-1)$-dimensional Lebesgue measure on $\mathbb{R}^{\{1,\ldots,r\}\setminus\{l\}}$ by the *lifting mapping* to $\mathsf{aff}(\Theta) \subseteq \mathbb{R}^{\{1,\ldots,r\}}$, for arbitrarily chosen $l \in \{1, \ldots, r\}$:

$$[\theta_k]_{k\in\{1,\ldots,r\}\setminus\{l\}} \quad \longmapsto \quad (\,[\theta_k]_{k\in\{1,\ldots,r\}\setminus\{l\}}\,,\, \theta_l \equiv 1 - \sum_{k=1, k\neq l}^{r} \theta_k\,) \tag{5}$$

It appears the image of the Lebesgue measure on $\mathbb{R}^{\{1,\ldots,r\}\setminus\{l\}}$ by this transformation does not depend on the choice of $l \in \{1, \ldots, r\}$ and, moreover, it is nothing but the $\frac{1}{\sqrt{r}}$-multiple of the proper Lebesgue measure on $\mathsf{aff}(\Theta)$ - see Proposition 17 in § B.4.

Definition 3 (dominating measure for Dirichlet distributions)
Given $\Theta \equiv \{\, \theta \equiv (\theta_1, \ldots, \theta_r)\,;\ \theta_k > 0,\ \sum_{k=1}^r \theta_k = 1 \,\}$, $r \geq 2$ the symbol μ_Θ will denote the restriction (to Θ) of the image of the $(r-1)$-dimensional Lebesgue measure by (5). It will serve as a *standard dominating measure for Dirichlet distributions.*

Remark The specification of the dominating measure (for Dirichlet distributions) is typically omitted in the literature. Indeed, I was not able to find that particular piece of information in any book dealing with Dirichlet distributions (I had a chance to consult). Perhaps some of the authors consider it to be intuitively clear that the lifting transformation (5) always leads to the same measure on $\mathsf{aff}\,(\Theta)$. However, since this transformation is not an isometry (neither a multiple of an isometry), the proof of this fact is not straightforward and deserves some special geometric considerations – see § B.4.

4.2 Definition of Dirichlet distributions

The class of Dirichlet distributions can be introduced as follows.

Fixed parameter: $r \geq 2$
This is the dimension of the Euclidean space in which the sample space for Dirichlet distributions is placed. Alternatively, the meaning of r is the number of outcomes (of the ascribed experiment). Formally, it could also be $r = 1$, but then the corresponding Dirichlet distribution is a degenerate discrete distribution.

Sample space: $\Theta = \{\, (\theta_1, \ldots, \theta_r)\,;\ \theta_k > 0,\ \sum_{k=1}^r \theta_k = 1 \,\}$
As mentioned above, this is the interior of the probability simplex in $\mathbb{R}^r$, which serves as the parameter space for (the classes of) discrete distributions introduced in § 3.

Parameter space: $\Xi = \{\, \alpha \equiv (\alpha_1, \ldots, \alpha_r)\,;\ \alpha_k > 0 \,\}$
This is the interior of the positive quadrant in $\mathbb{R}^r$. To distinguish terminologically this space Ξ from Θ, which usually plays the role of the parameter space, Ξ will be named the *hyper-parameter space* and its elements (vector) *hyper-parameters*. Observe that there is no functional dependence between hyper-parameter vector components (= single hyper-parameters). Therefore, the actual dimension of Ξ is higher than the dimension of Θ (by 1).

The formula for the density (with respect to μ_Θ from Definition 3):

$$\forall\, \alpha \equiv (\alpha_1, \ldots, \alpha_r) \in \Xi \quad \forall\, \theta \equiv (\theta_1, \ldots, \theta_r) \in \Theta$$
$$d_\alpha([\theta_1, \ldots, \theta_r]) \;=\; \frac{\Gamma(\sum_{k=1}^r \alpha_k)}{\prod_{k=1}^r \Gamma(\alpha_k)} \cdot \prod_{k=1}^r (\theta_k)^{\alpha_k - 1}\,. \tag{6}$$

The normalizing constant in (6) is the correct one because $\int_\Theta d_\alpha(\theta)\, d\mu_\Theta(\theta) = 1$ for each $\alpha \in \Xi$. This follows from the formula (10) presented in the next section.

The Dirichlet distribution corresponding to the collection of hyper-parameters $[\alpha_k]_{k=1}^r$ will be denoted by $\mathcal{D}([\alpha_k]_{k=1}^r)$.

4.3 Auxiliary formula

The following observation simplifies many computations concerning Dirichlet distributions.

Proposition 1 $\quad \forall\, r \geq 2,\ \ \forall\, \alpha_1, \ldots, \alpha_r > 0$

$$\int\limits_{\substack{\theta_1,\ldots,\theta_{r-1}>0 \\ \sum_{k=1}^{r-1}\theta_k<1}} \{\prod_{k=1}^{r-1}(\theta_k)^{\alpha_k-1}\}\cdot\{1-\sum_{k=1}^{r-1}\theta_k\}^{\alpha_r-1}\ d\theta_1\ldots\theta_{r-1} \;=\; \frac{\prod_{k=1}^{r}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{r}\alpha_k)}\,. \tag{7}$$

Proof. This can be shown by induction on $r \geq 2$. The induction hypothesis for $r = 2$ follows from well-known formulas for Beta and Gamma function (see § A):

$$\begin{aligned}\int\limits_{\substack{\theta_1>0 \\ \theta_1<1}} \theta_1^{\alpha_1-1}\cdot\{1-\theta_1\}^{\alpha_2-1}\,d\theta_1 \;&=\; \int_0^1 \theta^{\alpha_1-1}\cdot(1-\theta)^{\alpha_2-1}\,d\theta \\ &=\; B(\alpha_1,\alpha_2) = \frac{\Gamma(\alpha_1)\cdot\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)} = \frac{\prod_{k=1}^{2}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{2}\alpha_k)}\,.\end{aligned}$$

Now, we assume $r \geq 3$ and try verify the induction step. First, we observe that the induction premise implies the formula: $\forall\, \alpha_1, \ldots, \alpha_{r-1} > 0$

$$\int\limits_{\substack{\theta_1,\ldots,\theta_{r-2}>0 \\ \sum_{k=1}^{r-2}\theta_k<1}} \{\prod_{k=1}^{r-2}(\theta_k)^{\alpha_k-1}\}\cdot\{1-\sum_{k=1}^{r-2}\theta_k\}^{\alpha_{r-1}-1}\ d\theta_1\ldots\theta_{r-2} \;=\; \frac{\prod_{k=1}^{r-1}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{r-1}\alpha_k)}\,. \tag{8}$$

The next step is to verify that (8) implies a formally stronger version:
$\forall\, 0 < \gamma \leq 1, \quad \alpha_1, \ldots, \alpha_{r-1} > 0$

$$\int\limits_{\substack{\eta_1,\ldots,\eta_{r-2}>0 \\ \sum_{k=1}^{r-2}\eta_k<\gamma}} \{\prod_{k=1}^{r-2}(\eta_k)^{\alpha_k-1}\} \quad\cdot\quad \{\gamma-\sum_{k=1}^{r-2}\eta_k\}^{\alpha_{r-1}-1} \quad d\eta_1\ldots\eta_{r-2}$$

$$=\; \gamma^{(\sum_{k=1}^{r-1}\alpha_k)-1}\cdot\frac{\prod_{k=1}^{r-1}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{r-1}\alpha_k)}\,. \tag{9}$$

Indeed, to see (8)⇒(9) we apply the substitution $\int_{y\in\varphi(X)} f(y)\,d\mu\varphi^{-1}(y) = \int_{x\in X} f\circ\varphi(x)\,d\mu(x)$, where we specifically have $x \equiv [\theta_1, \ldots, \theta_{r-2}]$, X is determined by inequalities $\theta_1, \ldots, \theta_{r-2} > 0$, $\sum_{k=1}^{r-2}\theta_k < 1$ and $\varphi(x) = y \equiv [\eta_1, \ldots, \eta_{r-2}]$ is defined by the relation $\eta_k = \gamma\cdot\theta_k$ for $k = 1, \ldots, r-2$. Then $\varphi(X)$ is the domain of the integral in (9). The measure μ is the restriction (to X) of the product of γ-multiples of one-dimensional Lebesgue measures. Thus, $\mu\varphi^{-1}$ is the product of one-dimensional Lebesgue measures, that is, the $(r-2)$-dimensional Lebesgue measure. The function f is the argument of the integral in (9): $f(y) = \{\prod_{k=1}^{r-2}(\eta_k)^{\alpha_k-1}\}\cdot\{\gamma-\sum_{k=1}^{r-2}\eta_k\}^{\alpha_{r-1}-1}$. By substituting $\eta_k = \gamma\cdot\theta_k$ for $k = 1, \ldots, r-2$ to this we get the expression for $f\circ\varphi(x)$:

$$\begin{aligned} f\circ\varphi(x) \;&=\; \prod_{k=1}^{r-2}(\gamma\cdot\theta_k)^{\alpha_k-1}\cdot\{\gamma-\sum_{k=1}^{r-2}(\gamma\cdot\theta_k)\}^{\alpha_{r-1}-1} \\ &=\; \prod_{k=1}^{r-2}\gamma^{\alpha_k-1}\cdot\prod_{k=1}^{r-2}\theta_k^{\alpha_k-1}\cdot\{\gamma-\gamma\cdot\sum_{k=1}^{r-2}\theta_k\}^{\alpha_{r-1}-1} \\ &=\; \prod_{k=1}^{r-2}\gamma^{\alpha_k-1}\cdot\prod_{k=1}^{r-2}\theta_k^{\alpha_k-1}\cdot\gamma^{\alpha_{r-1}-1}\cdot\{1-\sum_{k=1}^{r-2}\theta_k\}^{\alpha_{r-1}-1} \\ &=\; \prod_{k=1}^{r-1}\gamma^{\alpha_k-1}\cdot\prod_{k=1}^{r-2}\theta_k^{\alpha_k-1}\cdot\{1-\sum_{k=1}^{r-2}\theta_k\}^{\alpha_{r-1}-1}\,. \end{aligned}$$

Thus, by using the substitution we express the left-hand side of (9) in the following form:

$$\begin{aligned}
\int\limits_{x\in X} f\circ\varphi(x)\,d\mu(x) &= \int\limits_{\substack{\theta_1,\dots,\theta_{r-2}>0\\ \sum_{k=1}^{r-2}\theta_k<1}} \prod_{k=1}^{r-1}\gamma^{\alpha_k-1}\cdot\prod_{k=1}^{r-2}\theta_k^{\alpha_k-1}\cdot\{1-\sum_{k=1}^{r-2}\theta_k\}^{\alpha_{r-1}-1}\cdot\gamma^{r-2}\;d\theta_1\dots\theta_{r-2}\\
&= (\prod_{k=1}^{r-1}\gamma^{\alpha_k-1})\cdot\gamma^{r-2}\cdot\int\limits_{\substack{\theta_1,\dots,\theta_{r-2}>0\\ \sum_{k=1}^{r-2}\theta_k<1}}\prod_{k=1}^{r-2}\theta_k^{\alpha_k-1}\cdot\{1-\sum_{k=1}^{r-2}\theta_k\}^{\alpha_{r-1}-1}\,d\theta_1\dots\theta_{r-2}\\
&= \gamma^{(\sum_{k=1}^{r-1}\alpha_k)-(r-1)+(r-2)}\cdot\frac{\prod_{k=1}^{r-1}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{r-1}\alpha_k)} = \gamma^{(\sum_{k=1}^{r-1}\alpha_k)-1}\cdot\frac{\prod_{k=1}^{r-1}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{r-1}\alpha_k)}\,.
\end{aligned}$$

Here, in the first line, we used the fact that μ is $(r-2)$-multiple product of γ-multiples of the (one-dimensional) Lebesgue measures; then (8) was used in the third line. This concludes the proof of the implication (8)⇒(9).
Now, the formula (9) can be used to verify the induction step by means of the Fubini theorem, where one has $\gamma = 1-\theta_{r-1}$:

$$\int\limits_{\substack{\theta_1,\dots,\theta_{r-1}>0\\ \sum_{k=1}^{r-1}\theta_k<1}} \{\prod_{k=1}^{r-1}(\theta_k)^{\alpha_k-1}\}\cdot\{1-\sum_{k=1}^{r-1}\theta_k\}^{\alpha_r-1}\;\;d\theta_1\dots\theta_{r-1}$$

$$= \int\limits_{0<\theta_{r-1}<1}\theta_{r-1}^{\alpha_{r-1}-1}\cdot\left[\int\limits_{\substack{\theta_1,\dots,\theta_{r-2}>0\\ \sum_{k=1}^{r-2}\theta_k<1-\theta_{r-1}}}\{\prod_{k=1}^{r-2}(\theta_k)^{\alpha_k-1}\}\cdot\{1-\theta_{r-1}-\sum_{k=1}^{r-2}\theta_k\}^{\alpha_r-1}\;\;d\theta_1\dots\theta_{r-2}\right]d\theta_{r-1}\,.$$

The internal integral has, by (9), where we replace α_{r-1} by α_r and put $\gamma = 1-\theta_{r-1}$, the value

$$(1-\theta_{r-1})^{(\sum_{k=1,k\neq r-1}^{r}\alpha_k)-1}\cdot\frac{\prod_{k=1,k\neq r-1}^{r}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1,k\neq r-1}^{r}\alpha_k)}\,,$$

and one can write, using well-known formulas for Beta and Gamma function from § A:

$$\begin{aligned}
&\int\limits_{0<\theta_{r-1}<1}\theta_{r-1}^{\alpha_{r-1}-1}\cdot(1-\theta_{r-1})^{(\sum_{k=1,k\neq r-1}^{r}\alpha_k)-1}\cdot\frac{\prod_{k=1,k\neq r-1}^{r}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1,k\neq r-1}^{r}\alpha_k)}\;\;d\theta_{r-1}\\
&= \frac{\prod_{k=1,k\neq r-1}^{r}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1,k\neq r-1}^{r}\alpha_k)}\cdot\int_0^1\theta^{\alpha_{r-1}-1}\cdot(1-\theta)^{(\sum_{k=1,k\neq r-1}^{r}\alpha_k)-1}\,d\theta\\
&= \frac{\prod_{k=1,k\neq r-1}^{r}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1,k\neq r-1}^{r}\alpha_k)}\cdot B(\alpha_{r-1},\sum_{k=1,k\neq r-1}^{r}\alpha_k)\\
&= \frac{\prod_{k=1,k\neq r-1}^{r}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1,k\neq r-1}^{r}\alpha_k)}\cdot\frac{\Gamma(\alpha_{r-1})\cdot\Gamma(\sum_{k=1,k\neq r-1}^{r}\alpha_k)}{\Gamma(\sum_{k=1}^{r}\alpha_k)} = \frac{\prod_{k=1}^{r}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{r}\alpha_k)}\,,
\end{aligned}$$

which concludes the induction step. □

The above formula has the following consequence

Corollary 2 *Given $r \geq 2$, denote $\alpha_+ \equiv \sum_{k=1}^r \alpha_k$ for any $\alpha_1, \ldots, \alpha_r > 0$. Then*

$$\int_\Theta \prod_{k=1}^r (\theta_k)^{\alpha_k - 1} \; d\mu_\Theta(\theta) \; = \; \frac{\prod_{k=1}^r \Gamma(\alpha_k)}{\Gamma(\alpha_+)} \, . \tag{10}$$

Proof. As explained in § 4.1, μ_Θ is the image of the Lebesgue measure on $\mathbb{R}^{\{1,\ldots,r-1\}}$ by the lifting mapping

$$\mathcal{L} : [\theta_k]_{k=1}^{r-1} \quad \longmapsto \quad ([\theta_k]_{k=1}^{r-1}, \theta_r \equiv 1 - \sum_{k=1}^{r-1} \theta_k) \, ,$$

which is the specification of (5) for $l = r$. Therefore, the integral after μ_Θ in the left-hand side of (10) can be computed by the substitution $\mathcal{L}$; this means one has to compute the integral (7) from Proposition 1. □

The above formula (10) implies easily the formulas for the expectation and covariance of the Dirichlet distribution: (c.f. page 269 in [9] or page 47 in [6]):

$$E(\theta_k) = \frac{\alpha_k}{\alpha_+}, \qquad var(\theta_k) = \frac{\alpha_k \cdot (\alpha_+ - \alpha_k)}{\alpha_+^2 \cdot (\alpha_+ + 1)} \quad \text{for } k \in \{1, \ldots r\},$$
$$cov(\theta_k, \theta_l) = \frac{-\alpha_k \cdot \alpha_l}{\alpha_+^2 \cdot (\alpha_+ + 1)} \quad \text{for } k \neq l, \; k, l \in \{1, \ldots r\}.$$

Indeed, one can write using the formula (6):

$$E(\theta_k) = \int_\Theta \theta_k \cdot \underbrace{\frac{\Gamma(\alpha_+)}{\prod_{n=1}^r \Gamma(\alpha_n)} \cdot \prod_{n=1}^r (\theta_n)^{\alpha_n - 1}}_{d_\alpha([\theta_1, \ldots, \theta_r])} \; d\mu_\Theta(\theta) \; = \; \frac{\Gamma(\alpha_+)}{\prod_{n=1}^r \Gamma(\alpha_n)} \cdot \int_\Theta \prod_{n=1}^r (\theta_n)^{\beta_n - 1} \; d\mu_\Theta(\theta) \, ,$$

where $\beta_k = \alpha_k + 1$ and $\beta_n = \alpha_n$ for $n \in \{1, \ldots r\} \setminus \{k\}$. Thus, using the formula (10) for a different system of hyper-parameters:

$$\begin{aligned} E(\theta_k) &= \frac{\Gamma(\alpha_+)}{\prod_{n=1}^r \Gamma(\alpha_n)} \cdot \frac{\prod_{n=1}^r \Gamma(\beta_n)}{\Gamma(\beta_+)} = \frac{\Gamma(\alpha_+)}{\Gamma(\alpha_k)} \cdot \frac{\Gamma(\beta_k)}{\Gamma(\beta_+)} = \frac{\Gamma(\alpha_+)}{\Gamma(\alpha_+ + 1)} \cdot \frac{\Gamma(\alpha_k + 1)}{\Gamma(\alpha_k)} \\ &= \frac{\Gamma(\alpha_+)}{\alpha_+ \cdot \Gamma(\alpha_+)} \cdot \frac{\alpha_k \cdot \Gamma(\alpha_k)}{\Gamma(\alpha_k)} = \frac{\alpha_k}{\alpha_+} \, , \end{aligned}$$

where we have used the well-known formula (32). Analogously, one can derive

$$E(\theta_k^2) = \frac{\Gamma(\alpha_+)}{\Gamma(\alpha_+ + 2)} \cdot \frac{\Gamma(\alpha_k + 2)}{\Gamma(\alpha_k)} = \frac{(\alpha_k + 1) \cdot \alpha_k}{(\alpha_+ + 1) \cdot \alpha_+} \, ,$$

and then we use the well-known formula for the variance:

$$var(\theta_k) = E(\theta_k^2) - E(\theta_k)^2 = \frac{(\alpha_k + 1) \cdot \alpha_k}{(\alpha_+ + 1) \cdot \alpha_+} - \frac{\alpha_k^2}{\alpha_+^2} = \ldots = \frac{\alpha_k \cdot (\alpha_+ - \alpha_k)}{\alpha_+^2 \cdot (\alpha_+ + 1)} \, .$$

By the same procedure

$$E(\theta_k \cdot \theta_l) = \frac{\Gamma(\alpha_+)}{\Gamma(\alpha_+ + 2)} \cdot \frac{\Gamma(\alpha_k + 1)}{\Gamma(\alpha_k)} \cdot \frac{\Gamma(\alpha_l + 1)}{\Gamma(\alpha_l)} = \frac{\alpha_k \cdot \alpha_l}{(\alpha_+ + 1) \cdot \alpha_+} \, ,$$

and by the well-known formula for the covariance:

$$cov(\theta_k, \theta_l) = E(\theta_k \cdot \theta_l) - E(\theta_k) \cdot E(\theta_l) = \frac{\alpha_k \cdot \alpha_l}{(\alpha_+ + 1) \cdot \alpha_+} - \frac{\alpha_k}{\alpha_+} \cdot \frac{\alpha_l}{\alpha_+} = \ldots = \frac{-\alpha_k \cdot \alpha_l}{\alpha_+^2 \cdot (\alpha_+ + 1)} \, .$$

Remark The consequence of the formulas above is that, for $r \geq 2$, the mapping $[\alpha_k]_{k=1}^r \longmapsto \mathcal{D}([\alpha_k]_{k=1}^r)$ is an injective mapping.

Indeed, let $\alpha = [\alpha_k]_{k=1}^r$ and $\beta = [\beta_k]_{k=1}^r$ be two strictly positive vectors and assume $\mathcal{D}([\alpha_k]_{k=1}^r) = \mathcal{D}([\beta_k]_{k=1}^r)$. Then $\forall\, k$ one has $\frac{\alpha_k}{\alpha_+} = E(\theta_k) = \frac{\beta_k}{\beta_+}$, which means $\beta_k = t \cdot \alpha_k$, where $t \equiv \frac{\beta_+}{\alpha_+} > 0$. Hence, by the formula for the variance and the substitution $\beta_+ = t \cdot \alpha_+$, $\beta_k = t \cdot \alpha_k$:

$$\frac{\alpha_k \cdot (\alpha_+ - \alpha_k)}{\alpha_+^2 \cdot (\alpha_+ + 1)} = var(\theta_k) = \frac{\beta_k \cdot (\beta_+ - \beta_k)}{\beta_+^2 \cdot (\beta_+ + 1)} = \frac{t \cdot \alpha_k \cdot (t \cdot \alpha_+ - t \cdot \alpha_k)}{t^2 \cdot \alpha_+^2 \cdot (t \cdot \alpha_+ + 1)} = \frac{\alpha_k \cdot (\alpha_+ - \alpha_k)}{\alpha_+^2 \cdot (t \cdot \alpha_+ + 1)}\,.$$

Hence, by canceling, $t \cdot \alpha_+ + 1 = \alpha_+ + 1$. Thus, $t = 1$ and $\beta_k = \alpha_k$ for every k.

Remark (the relation of the Dirichlet distribution and beta distribution)
A common distribution in statistics is *beta distribution*, defined as follows: the parameters are $a, b > 0$, the sample space is the interval $(0, 1)$ and the density (with respect to the Lebesgue measure) is given by the formula

$$f(x) = \frac{1}{B(a,b)} \cdot x^{a-1} \cdot (1-x)^{b-1}\,,$$

where $B(a, b)$ is the value of the Beta function. Note that $\frac{1}{B(a,b)}$ is the correct normalizing constant (see § A) and $a = b = 1$ gives the uniform distribution on $(0, 1)$.

Beta distribution is, in fact, one-dimensional marginal of the Dirichlet distribution.[7] More specifically, if $r = 2$ the Dirichlet distribution is settled on a slantwise segment $\Theta = \{(\theta_1, \theta_2);\, \theta_1, \theta_2 > 0,\ \theta_1 + \theta_2 = 1\}$. For parameters $\alpha_1, \alpha_2 > 0$, its density is given by

$$d_\alpha(\theta_1, \theta_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2)} \cdot (\theta_1)^{\alpha_1 - 1} \cdot (\theta_2)^{\alpha_2 - 1}\,.$$

This is meant with respect to the image of the Lebesgue measure on $(0, 1)$ by the lifting mapping $\mathcal{L}_1 : \theta_1 \mapsto (\theta_1, 1 - \theta_1)$, or alternatively, by the mapping $\mathcal{L}_2 : \theta_2 \mapsto (1 - \theta_2, \theta_2)$. Thus, $\mathcal{D}(\alpha_1, \alpha_2)$ can be interpreted as the beta distribution $\beta(\alpha_1, \alpha_2)$ transformed by $\mathcal{L}_1$ to Θ, or alternatively, as the beta distribution $\beta(\alpha_2, \alpha_1)$ transformed by $\mathcal{L}_2$. It is evident, that the marginal of $\mathcal{D}(\alpha_1, \alpha_2)$ for θ_1 is $\beta(\alpha_1, \alpha_2)$ and its marginal for θ_2 is $\beta(\alpha_2, \alpha_1)$.

Finally, it is easy to see that the class of Dirichlet distributions is also an exponential family (for fixed $r \geq 2$):

- μ is the dominating measure μ_Θ from Definition 3 (in § 4.1),
- $m = r$,
- $c(\alpha) \equiv \dfrac{\Gamma(\sum_{k=1}^r \alpha_k)}{\prod_{k=1}^r \Gamma(\alpha_k)}$ for $\alpha \in \Xi$,
- $u(\theta) = 1$ for $\theta \in \Theta$,
- $q_s(\alpha) = \alpha_s - 1$ for $\alpha \in \Xi$, $s = 1, \ldots, r$,
- $t_s(\theta) = \ln \theta_s$ for $\theta \in \Theta$, $s = 1, \ldots, r$.

[7] Note that the class of Dirichlet distributions is not closed under marginalization.

To check that let us substitute to the formula (1):

$$\begin{aligned} d_\alpha(\theta) &= c(\alpha)\cdot u(\theta)\cdot \exp\big(\sum_{s=1}^{m} q_s(\alpha)\cdot t_s(\theta)\big) = \frac{\Gamma(\sum_{k=1}^r \alpha_k)}{\prod_{k=1}^r \Gamma(\alpha_k)}\cdot 1\cdot \exp\big(\sum_{s=1}^{r}(\alpha_s-1)\cdot \ln\theta_s\big) \\ &= \frac{\Gamma(\sum_{k=1}^r \alpha_k)}{\prod_{k=1}^r \Gamma(\alpha_k)}\cdot \exp\big(\sum_{s=1}^{r}\ln(\theta_s)^{\alpha_s-1}\big) = \frac{\Gamma(\sum_{k=1}^r \alpha_k)}{\prod_{k=1}^r \Gamma(\alpha_k)}\cdot \prod_{s=1}^{r}\exp\ln(\theta_s)^{\alpha_s-1} \\ &= \frac{\Gamma(\sum_{k=1}^r \alpha_k)}{\prod_{k=1}^r \Gamma(\alpha_k)}\cdot \prod_{s=1}^{r}(\theta_s)^{\alpha_s-1} = \frac{\Gamma(\sum_{k=1}^r \alpha_k)}{\prod_{k=1}^r \Gamma(\alpha_k)}\cdot \prod_{k=1}^{r}(\theta_k)^{\alpha_k-1}, \end{aligned}$$

which is the expression in (6).

5 Bayesian approach

The aim of this section is recall the basic idea of the Bayesian approach. This approach will be later applied, in § 7 and § 8, to the special case of a Bayesian network model.

5.1 Bayesian terminology

First, following the Preface of the book [12], we recapitulate the basic terminology in this area. A so-called Bayesian experiment is specified by the following items:

- A measurable space $(\Theta, \mathcal{A})$, called the *parameter space*.
- A measurable space $(\mathbb{X}, \mathcal{X})$, called the *sample space*.
- A system $\mathcal{P} = \{P_\theta;\, \theta \in \Theta\}$ of probability distributions on $(\mathbb{X}, \mathcal{X})$ which is a Markov kernel from $(\Theta, \mathcal{A})$ to $(\mathbb{X}, \mathcal{X})$ (see § 2.2). They are called *sampling probabilities*.
  That means, the statistical model in the sense of Definition 1 is given.
  Usually, $\mathcal{P}$ is specified by a system of densities $p(x|\theta)$, $x \in \mathbb{X}$, $\theta \in \Theta$ with respect to a dominating measure on $(\mathbb{X}, \mathcal{X})$.
  Quite often, it is an exponential family in the sense of Definition 2.
- A probability distribution on $(\Theta, \mathcal{A})$, called the *prior probability*. It is usually defined by means of the density $\pi(\theta)$, $\theta \in \Theta$ with respect to a dominating measure on $(\Theta, \mathcal{A})$. The density is then called the *prior density*.

The components above define a probability measure $\mathbf{\Pi}$ on the product measurable space $(\Theta \times \mathbb{X}, \mathcal{A} \times \mathcal{X})$. This measure is called a "*Bayesian experiment*" by Florens, Mouchart and Rolin [12]. Typically, there exists a dual decomposition of $\mathbf{\Pi}$ to its marginal on $(\mathbb{X}, \mathcal{X})$ and the system probability distributions on $(\Theta, \mathcal{A})$ which is a Markov kernel from $(\mathbb{X}, \mathcal{X})$ to $(\Theta, \mathcal{A})$.[8] That means, the (joint) density $\Pi(\theta, x)$ of $\mathbf{\Pi}$ (with respect to the product of "standard" dominating measures on $(\Theta, \mathcal{A})$ and $(\mathbb{X}, \mathcal{X})$) and can be written as follows:

$$\pi(\theta)\cdot p(x|\theta) \equiv \Pi(\theta, x) = p(x)\cdot \pi(\theta|x) \qquad \text{for } \theta \in \Theta,\ x \in \mathbb{X}\,.$$

This allows one to introduce two other important components of the Bayesian experiment:

[8] This is ensured, for example, if sampling probabilities are given by a system of densities with respect to a dominating measure, or alternatively, by suitable topological assumptions on the parameter space $(\Theta, \mathcal{A})$.

- The marginal measure of $\mathbf{\Pi}$ on $(\mathbb{X}, \mathcal{X})$ is called the *predictive probability*. Its density $p(x)$ is obtained by integrating the joint density $\Pi(\theta, x)$ over the space $(\Theta, \mathcal{A})$ (with respect to the corresponding dominating measure).

  Some authors [9], in the case $\mathbb{X}$ is the collection of all samples of the size d (see § 3.1), call this marginal measure the "*marginal probability of data*".

- The system of probability measures $\{\boldsymbol{\pi}_x;\, x \in \mathbb{X}\}$ will be called the system of *posterior probabilities*. Their densities $\pi(\theta|x)$ can be obtained by dividing the joint density $\Pi(\theta, x)$ by the density $p(x)$ of the predictive probability.

The Bayesian approach consists in the interpretation of the above mentioned components. For example, if $x \in \mathbb{X}$ is interpreted as the outcome of a series of measurements then the posterior density $\pi_x(\theta) \equiv \pi(\theta|x)$ describes the adaptation of the original (probabilistic) knowledge about the parameters expressed in the form of the prior density $\pi(\theta)$.

5.2 Conjugate family

Quite often, not just a single prior distribution, but a whole (parameterized) class of distributions $\mathcal{S}$ on the parameter space $(\Theta, \mathcal{A})$ is considered.[9] Prior distributions are then chosen from this class $\mathcal{S}$. The potential parameters of probability measures in $\mathcal{S}$ are then called *hyper-parameters*. The usual technical requirement is as follows.

Definition 4 (conjugate family)
A class $\mathcal{S}$ of probability distributions on the parameter space $(\Theta, \mathcal{A})$ will be called *conjugate* to a class $\mathcal{T}$ of probability distributions on the sample space $(\mathbb{X}, \mathcal{X})$ if the following condition is valid: whenever the prior probability belongs to $\mathcal{S}$ and the sampling probabilities to $\mathcal{T}$ then every posterior probability belongs to $\mathcal{S}$, that is,

$$\pi(\theta) \in \mathcal{S}, \quad \forall\, \theta \in \Theta \;\; p(x|\theta) \in \mathcal{T} \quad \Rightarrow \quad \forall\, x \in \mathbb{X} \;\; \pi(\theta|x) \in \mathcal{S}\,.$$

The importance of the class of Dirichlet distributions (see § 4) consists in the fact it is a conjugate family to (each of) the discrete statistical models from § 3. Below we show this both for the class of sampling distributions from § 3.1 and the class of multinomial distributions from § 3.2.[10]

5.2.1 Conjugacy of Dirichlet distributions to sampling distributions

Assume $d, r \in \mathbb{Z}$, $d \geq 1$, $r \geq 2$ are fixed. Consider the following spaces (cf. § 4.2 and § 3.1) :

- $\Xi = \{\, \alpha \equiv (\alpha_1, \ldots, \alpha_r)\,;\; \alpha_k > 0 \,\}$ is the set of hyper-parameters,
- $\Theta = \{\, \theta \equiv (\theta_1, \ldots, \theta_r)\,;\; \theta_k > 0,\; \sum_{k=1}^{r} \theta_k = 1 \,\}$ is the parameter space endowed with the dominating measure μ_Θ (see Definition 3 in § 4.1), and
- $\mathbb{X} = X^{\{1,\ldots,d\}}$, where $X = \{a_1, \ldots a_r\}$, is the sample space. The dominating measure on $\mathbb{X}$ is the arithmetic measure.

[9] Again, it is often an exponential family in the sense of Definition 2.

[10] The statistical model for the theoretical distributions (= a single outcome) mentioned in the beginning of § 3 is a special case of the model for sampling distributions from § 3.1: one has $d = 1$ in this case.

Now, assuming $\alpha \in \Xi$ is fixed, (6) gives the formula for the respective prior density:

$$\forall\, \theta \in \Theta \qquad \pi(\theta) \equiv d_\alpha(\theta) = \frac{\Gamma(\sum_{k=1}^r \alpha_k)}{\prod_{k=1}^r \Gamma(\alpha_k)} \cdot \prod_{k=1}^r \, (\theta_k)^{\alpha_k - 1} \,.$$

Analogously, (2) gives the formula for sampling densities:

$$\forall\, y \equiv [y^1, \ldots, y^d] \in X^{\{1,\ldots,d\}}, \quad \forall\, \theta \in \Theta$$
$$p(y|\theta) = \prod_{k=1}^r (\theta_k)^{x_k} \qquad \text{where} \;\; x_k \equiv |\{ y^\ell;\, 1 \le \ell \le d,\; y^\ell = a_k \}| \;\; \text{for } k = 1, \ldots, r \,.$$

Therefore, the joint density has the form

$$\Pi(\theta, y) = \pi(\theta) \cdot p(y|\theta) = \frac{\Gamma(\sum_{k=1}^r \alpha_k)}{\prod_{k=1}^r \Gamma(\alpha_k)} \cdot \prod_{k=1}^r \, (\theta_k)^{\alpha_k - 1} \cdot \prod_{k=1}^r (\theta_k)^{x_k} = \frac{\Gamma(\sum_{k=1}^r \alpha_k)}{\prod_{k=1}^r \Gamma(\alpha_k)} \cdot \prod_{k=1}^r \, (\theta_k)^{\alpha_k + x_k - 1} \,.$$

Hence, one can get the predictive density by integrating: $\forall\, y \in X^{\{1,\ldots,d\}}$

$$\begin{aligned} p(y) \;\; &= \;\; \int_\Theta \, \Pi(\theta, y) \;\, d\mu_\Theta(\theta) \;=\; \int_\Theta \, \frac{\Gamma(\sum_{k=1}^r \alpha_k)}{\prod_{k=1}^r \Gamma(\alpha_k)} \cdot \prod_{k=1}^r \, (\theta_k)^{\alpha_k + x_k - 1} \;\, d\mu_\Theta(\theta) \\ &= \;\; \frac{\Gamma(\sum_{k=1}^r \alpha_k)}{\prod_{k=1}^r \Gamma(\alpha_k)} \cdot \int_\Theta \, \prod_{k=1}^r \, (\theta_k)^{\alpha_k + x_k - 1} \;\, d\mu_\Theta(\theta) \;=\; \frac{\Gamma(\sum_{k=1}^r \alpha_k)}{\prod_{k=1}^r \Gamma(\alpha_k)} \cdot \frac{\prod_{k=1}^r \Gamma(\alpha_k + x_k)}{\Gamma(\sum_{k=1}^r \alpha_k + x_k)} \,, \end{aligned}$$

where in the second line we used the formula (10) from Corollary 2 in § 4.3.

Thus, the formula for the predictive density in terms of $\alpha \in \Xi$ is as follows:[11]

$$p_\alpha(y) = \frac{\Gamma(\sum_{k=1}^r \alpha_k)}{\prod_{k=1}^r \Gamma(\alpha_k)} \cdot \frac{\prod_{k=1}^r \Gamma(\alpha_k + x_k)}{\Gamma(\sum_{k=1}^r \alpha_k + x_k)} \qquad \text{where} \;\; x_k \equiv |\{ y^\ell;\, 1 \le \ell \le d,\; y^\ell = a_k \}| \,. \quad (11)$$

Having fixed $y \in X^{\{1,\ldots,d\}}$ one can compute the posterior density by dividing:

$$\pi(\theta|y) = \frac{\Pi(\theta, y)}{p(y)} = \frac{\frac{\Gamma(\sum_{k=1}^r \alpha_k)}{\prod_{k=1}^r \Gamma(\alpha_k)} \cdot \prod_{k=1}^r \, (\theta_k)^{\alpha_k + x_k - 1}}{\frac{\Gamma(\sum_{k=1}^r \alpha_k)}{\prod_{k=1}^r \Gamma(\alpha_k)} \cdot \frac{\prod_{k=1}^r \Gamma(\alpha_k + x_k)}{\Gamma(\sum_{k=1}^r \alpha_k + x_k)}} = \frac{\Gamma(\sum_{k=1}^r \alpha_k + x_k)}{\prod_{k=1}^r \Gamma(\alpha_k + x_k)} \cdot \prod_{k=1}^r \, (\theta_k)^{\alpha_k + x_k - 1} \,.$$

That means, the posterior distribution is again Dirichlet; more specifically

$$\pi_\alpha(\theta|y) = \frac{\Gamma(\sum_{k=1}^r \alpha_k + x_k)}{\prod_{k=1}^r \Gamma(\alpha_k + x_k)} \cdot \prod_{k=1}^r \, (\theta_k)^{\alpha_k + x_k - 1} \equiv d_{\alpha + x}(\theta) \qquad (12)$$

where the vector $x \equiv [x_1, \ldots, x_r]$ is given by $x_k \equiv |\{ y^\ell;\, 1 \le \ell \le d,\; y^\ell = a_k \}|$ for $k = 1, \ldots, r$.

5.2.2 Conjugacy of Dirichlet distributions to multinomial distributions

Again, having $d, r \in \mathbb{Z}$, $d \ge 1$, $r \ge 2$ fixed, consider the spaces (cf. § 4.2 and § 3.2):

- the set of hyper-parameters $\Xi = \{\, \alpha \equiv (\alpha_1, \ldots, \alpha_r)\,;\; \alpha_k > 0 \,\}$,
- the parameter space $\Theta = \{\, \theta \equiv (\theta_1, \ldots, \theta_r)\,;\; \theta_k > 0,\; \sum_{k=1}^r \theta_k = 1 \,\}$ endowed with the dominating measure μ_Θ from Definition 3 (see § 4.1), and

[11] Observe that $p_\alpha(y) > 0$ for any $y \in \mathbb{X}$. Note also that (11) is nothing but the the formula (A.14) in [9].

- the sample space $\mathbb{X} = \{\, [x_1, \ldots, x_r]\,;\ x_k \in \{0, \ldots, d\}\ \ \sum_{k=1}^{r} x_k = d\} \subseteq \mathbb{R}^r$ with the arithmetic measure as the dominating measure.

Having $\alpha \in \Xi$ is fixed, the formula for the prior density (6) is the same as in § 5.2.1:

$$\forall\, \theta \in \Theta \qquad \pi(\theta) \equiv d_\alpha(\theta) = \frac{\Gamma(\sum_{k=1}^{r} \alpha_k)}{\prod_{k=1}^{r} \Gamma(\alpha_k)} \cdot \prod_{k=1}^{r} (\theta_k)^{\alpha_k - 1}\,,$$

but what is different is the formula for the sampling probabilities, now given by (4):

$$\forall\, x \equiv [x_1, \ldots, x_r] \in \mathbb{X},\ \forall\, \theta \equiv (\theta_1, \ldots, \theta_r) \in \Theta \quad p(x|\theta) = \frac{d!}{\prod_{k=1}^{r}(x_k!)} \cdot \prod_{k=1}^{r} (\theta_k)^{x_k}\,.$$

The only substantial difference is the additional factor, namely the multinomial coefficient $\frac{d!}{\prod_{k=1}^{r}(x_k!)}$. Thus, one can, more or less, repeat the calculations from § 5.2.1 and obtain the formula for the predictive density:

$$p_\alpha(x) = \frac{d!}{\prod_{k=1}^{r}(x_k!)} \cdot \frac{\Gamma(\sum_{k=1}^{r} \alpha_k)}{\prod_{k=1}^{r} \Gamma(\alpha_k)} \cdot \frac{\prod_{k=1}^{r} \Gamma(\alpha_k + x_k)}{\Gamma(\sum_{k=1}^{r} \alpha_k + x_k)} \quad \text{where } x \equiv [x_1, \ldots, x_r]\,. \tag{13}$$

Since the joint density has the same additional factor, in the calculation of the posterior density this additional factor cancels and the result will be the same formula as in § 5.2.1:

$$\pi_\alpha(\theta|x) = \frac{\Gamma(\sum_{k=1}^{r} \alpha_k + x_k)}{\prod_{k=1}^{r} \Gamma(\alpha_k + x_k)} \cdot \prod_{k=1}^{r} (\theta_k)^{\alpha_k + x_k - 1} \equiv d_{\alpha + x}(\theta)\,. \tag{14}$$

Thus, in both cases the "move" from the prior probability to the posterior probability is the change $\mathcal{D}([\alpha_k]_{k=1}^r) \longmapsto \mathcal{D}([\alpha_k + x_k]_{k=1}^r)$.

Remark The relation of predictive probabilities in these two cases is as follows. In the case of sampling distributions, the predictive probability (11) is the same for those samples (= databases) which yield the same table of counts. That means, it is constant on respective equivalence classes.[12] The predictive probability in the multinomial case is obtained from it by simple summing probabilities of samples yielding the same table of counts. Since they are equiprobable this reduces to the multiplication by the number of samples in the equivalence class, which is the corresponding multinomial coefficient, which gives (13).

6 Statistical model of a discrete Bayesian network

The theoretical basis of the procedures for learning a BN structure (by the method of maximizing a quality criterion) is the interpretation of every BN structure as a statistical model. This statistical model is specified in this section.

Throughout this section we fix an *acyclic directed graph* G over a finite non-empty set of nodes (= *variables*) N. Moreover, for every variable $i \in N$, we fix the *set* X_i *of possible values* for i.

[12] Here, two samples are considered to be equivalent if they yield the same table of counts.

6.1 Sample space and notational convention

We assume $|\mathsf{X}_i| \geq 2$ to ensure that the random variable corresponding to i is not degenerate and, therefore, at least one free parameter corresponds to i in the parameterization described below.

Assumption: $1 \leq |N| < \infty, \quad \forall\, i \in N \quad 2 \leq |\mathsf{X}_i| < \infty$

Sample space: The joint sample space will be the Cartesian product $\mathsf{X}_N \equiv \prod_{i \in N} \mathsf{X}_i$.

The following series of notational conventions is desirable to have elegant formulas for (Bayesian) quality criteria.

Conventions I
First, for every $A \subseteq N$, we denote by X_A the set of all possible *configurations for* A, that is, the set of lists $[x_i]_{i \in A}$ such that $x_i \in \mathsf{X}_i$ for $i \in A$.[13]

For every $A \subseteq N$ we *choose and fix an ordering of the set of configurations* X_A for A.

- The letter i will be used as *a generic symbol for a node/variable*: $\boxed{i \in N}$.
  Then $r(i)$ will denote the number of elements of X_i: $r(i) \equiv |\mathsf{X}_i|$ for $i \in N$.

- The letter k will be used as *a generic symbol for (codes of) node configurations*: $\boxed{k \in \{1, \ldots, r(i)\}}$.
  More specifically, for any node/variable $i \in N$ consider the chosen fixed ordering of configurations for $A = \{i\}$: $\mathsf{X}_i = \{y_i^1, y_i^2, \ldots, y_i^{r(i)}\}$. Then each $k \in \{1, \ldots, r(i)\}$ will be the code for y_i^k.[14] Moreover, for every $x \in \mathsf{X}_N$, the symbol $k(i, x)$ will denote the code of its marginal node configuration for i, that is, it is the unique $k \in \{1, \ldots, r(i)\}$ for which $x_i = y_i^k$.

Further conventions are related to the given acyclic directed graph G (over N). For every node/variable $i \in N$ let $pa_G(i)$ denote the set of *parents of* i *in* G:

$$pa_G(i) = \{h \in N\,;\ h \to i \text{ in } G\}\,.$$

Denote by $q(i, G)$ the number of *parent configurations* for i, that is, the number of elements of $\mathsf{X}_{pa_G(i)}$: $q(i, G) \equiv |\mathsf{X}_{pa_G(i)}|$ for $i \in N$.[15]

- The letter j will be used as *a generic symbol for (codes of) parent configurations*: $\boxed{j \in \{1, \ldots, q(i, G)\}}$.
  More specifically, for any node/variable $i \in N$ consider the fixed ordering of configurations for $A = pa_G(i)$: $\mathsf{X}_{pa_G(i)} = \{z_i^1, \ldots, z_i^{q(i,G)}\}$. Then $j \in \{1, \ldots, q(i, G)\}$ will be

[13] Observe that this definition of a configuration has reasonable sense for $A = \emptyset$. The configuration for the empty set is then the empty list. In particular, there exists (just one) configuration for the empty set, that is, $|\mathsf{X}_\emptyset| = 1$. On the other hand, if $A \neq \emptyset$ then X_A is nothing but the Cartesian product $\prod_{i \in A} \mathsf{X}_i$.

[14] In fact, a fully correct approach would be to consider for each $i \in N$ a distinct generic symbol k_i for the codes of elements of X_i. However, this would unnecessarily complicate later notation by second order indices. We can drop these superfluous indices because in the formulas we are going to write the generic symbol k will **always** be used in the situation the variable $i \in N$ is specified and there is no danger of misunderstanding.

[15] Recall that if $pa_G(i) = \emptyset$ then, by our definition of a configuration, $|\mathsf{X}_{pa_G(i)}| = |\mathsf{X}_\emptyset| = 1$ and, therefore, $q(i, G) = 1$ then.

the code for z_i^j, the j-th configuration in the ordering.[16] Moreover, for every $x \in \mathsf{X}_N$, the symbol $j(i,x)$ will denote the code of its marginal parent configuration for i, that is, the unique $j \in \{1,\ldots,q(i,G)\}$ such that $x_{pa_G(i)} = z_i^j$.

6.2 Parameter space

Now, it is possible to specify the parameter space. Single (one-dimensional) parameters correspond to (ordered) triplets

$$[\,\text{node} = \text{variable, parent configuration, node configuration}\,]\,,$$

where the configurations correspond to the variable. This is reflected by the notation for these single parameters with three indices:

$$\theta_{ijk} \quad \text{where } i \in N,\ j \in \{1,\ldots,q(i,G)\},\ k \in \{1,\ldots,r(i)\}\,.$$

The statistical model we consider here is the class of *strictly positive* probability distributions on X_N that are *Markovian* with respect to G.

Parameter space is the Cartesian product of (interiors of) probability simplices:

$$\Theta_G = \prod_{i\in N}\prod_{j=1}^{q(i,G)} \Theta_{(ij)} \quad \text{where } \Theta_{(ij)} \equiv \{\,[\theta_{ijk}]_{k=1}^{r(i)}\,;\ \theta_{ijk} > 0\ \ \sum_{k=1}^{r(i)} \theta_{ijk} = 1\,\}\,.$$

In particular, every vector parameter $\boldsymbol{\theta} \in \Theta_G$ decomposes into components:

$$\boldsymbol{\theta} = [\theta_{ijk}]_{\,i\in N,\, j\in\{1,\ldots,q(i,G)\},\, k\in\{1,\ldots,r(i)\}}\ .$$

Formula for the density (with respect to the arithmetic measure in X_N):

$$\forall\,\boldsymbol{\theta} \in \Theta_G \qquad p^{\boldsymbol{\theta}}(x) = \prod_{i\in N} \theta_{i\ j(i,x)\ k(i,x)} \quad \text{for } x \in \mathsf{X}_N. \tag{15}$$

One can verify (see Lemma 8.1 and Remark 8.4 in [27]) the following observations:

Lemma 3 *For every $\boldsymbol{\theta} \in \Theta_G$, $p^{\boldsymbol{\theta}}$ is the density of a strictly positive probability distribution on X_N. Moreover, the mapping $\boldsymbol{\theta} \mapsto p^{\boldsymbol{\theta}}$ is a one-to-one mapping of Θ_G onto the class $\mathcal{M}_+(G,\mathsf{X}_N)$ of strictly positive distributions on X_N that are Markovian with respect to G.*

Interpretation Moreover, it is shown in Lemma 8.1 from [27] that

$$\forall\, i \in N \quad \forall\, j \in \{1,\ldots,q(i,G)\} \quad \forall\, k \in \{1,\ldots,r(i)\}$$
$$\theta_{ijk} \text{ is the conditional probability } p^{\boldsymbol{\theta}}_{i|pa_G(i)}(y_i^k|z_i^j)\,.$$

In other words, every single parameter θ_{ijk} in the parameterization has the interpretation of the value of the conditional probability of the (corresponding) node configuration given the (corresponding) parent configuration.

[16] Again, it would be more precise to use, for each $i \in N$, a distinct generic symbol j_i for the codes of elements of $\mathsf{X}_{pa_G(i)}$. Nevertheless, like in the case of node configurations, the generic symbol j for parent configurations will **only** be used when the variable $i \in N$ is specified.

6.3 Exponential family

Having fixed the acyclic directed graph G (over N) and the sample space X_N, the above described parameterized class of Markovian distributions $\mathcal{M}_+(G, \mathsf{X}_N)$ defines a curved exponential family, which is a result reported already in [13]. To show this fact one needs to re-write the density (15) in the form (1):

$$p^{\boldsymbol{\theta}}(x) = c(\boldsymbol{\theta}) \cdot u(x) \cdot \exp\,(\,\sum_{s=1}^{m} t_s(x) \cdot q_s(\boldsymbol{\theta})\,) \qquad \text{for any } \boldsymbol{\theta} \in \Theta_G \text{ and } x \in \mathsf{X}_N\,.$$

More specifically, one has

- $m = \sum_{i\in N} \sum_{j=1}^{q(i,G)} \sum_{k=1}^{r(i)} 1 = \sum_{i\in N} r(i) \cdot q(i,G)$,
- $c(\boldsymbol{\theta}) \equiv 1$ for any $\boldsymbol{\theta} \in \Theta_G$,
- $u(x) = 1$ for every $x \in \mathsf{X}_N$,
- $q_s(\boldsymbol{\theta}) = \ln \theta_{ijk}$ for $s \sim (i,j,k)$,
- $t_s(x) = \delta(i,j,k|x)$ for $s \sim (i,j,k)$, where

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

  for any $i \in N$, $j \in \{1,\ldots,q(i,G)\}$, $k \in \{1,\ldots,r(i)\}$ and $x \in \mathsf{X}_N$.[17]

To evidence that let us re-write the formula (15):

$$\begin{aligned} p^{\boldsymbol{\theta}}(x) &= \prod_{i\in N} \theta_{i\,j(i,x)\,k(i,x)} = \prod_{i\in N} \prod_{j=1}^{q(i,G)} \prod_{k=1}^{r(i)} \theta_{ijk}^{\delta(i,j,k|x)} = \exp\,(\,\ln\,(\prod_i \prod_j \prod_k \theta_{ijk}^{\delta(i,j,k|x)}\,)\,) \\ &= \exp\,(\sum_i \sum_j \sum_k \ln\,\theta_{ijk}^{\delta(i,j,k|x)}\,) = 1 \cdot 1 \cdot \exp\,(\sum_{i\in N} \sum_{j=1}^{q(i,G)} \sum_{k=1}^{r(i)} \underbrace{\delta(i,j,k|x)}_{t_{ijk}(x)} \cdot \underbrace{\ln\,\theta_{ijk}}_{q_{ijk}(\boldsymbol{\theta})}\,) \\ &= c(\boldsymbol{\theta}) \cdot u(x) \cdot \exp\,(\,\sum_{s=1}^{m} t_s(x) \cdot q_s(\boldsymbol{\theta})\,), \end{aligned}$$

which is the required expression. Here, the second equality can be justified as follows: for fixed $i \in N$ the only θ_{ijk} for which $\delta(i,j,k|x) \neq 0$ is just $\theta_{i\,j(i,x)\,k(i,x)}$. Thus, if $\delta(i,j,k|x) = 1$ then $\theta_{ijk}^{\delta(i,j,k|x)} = \theta_{i\,j(i,x)\,k(i,x)}$, while if $\delta(i,j,k|x) = 0$ then $\theta_{ijk}^{\delta(i,j,k|x)} = 1$.

The actual dimension of Θ_G is $\sum_{i\in N}[r(i)-1]\cdot q(i,G)$, which is less than the number m above. That is why we only claim is a curved exponential family.

6.4 Likelihood function

The next step is to derive the formula for the likelihood function, that is, the density of the *sampling distribution*. To write an elegant expression for it one needs another series of conventions.

[17] By Conventions I, $x_{pa_G(i)} = z_i^j$ and $x_i = y_i^k$ is nothing but the requirement $j = j(i,x)$ and $k = k(i,x)$.

Conventions II
Let $D : x^1, \ldots, x^d$, $d \geq 1$ be an ordered sequence of elements of the joint sample space X_N, that is, a *database over N of the length d* ($\equiv$ a sample of the size d – see **B** in § 2.1).
If X_N is fixed then the symbol $\mathsf{DATA}(N, d)$ will denote the collection of all such databases.
Given $x \in \mathsf{X}_N$ and $D \in \mathsf{DATA}(N, d)$, $d \geq 1$ the symbol $d_{[x]}$ will denote the number of occurrences of x in the database:

$$d_{[x]} = |\{1 \leq \ell \leq d;\ x^\ell = x\}|\,.$$

Moreover, for every triplet (i, j, k) where $i \in N$, $j \in \{1, \ldots, q(i, G)\}$ and $k \in \{1, \ldots, r(i)\}$ the symbol d_{ijk} will denote the number of occurrences of the respective (marginal) configuration in D, analogously for a pair (i, j) where $i \in N$, $j \in \{1, \ldots, q(i, G)\}$:

$$\begin{aligned} d_{ijk} &= |\{1 \leq \ell \leq d;\ x^\ell_{\{i\}\cup \mathrm{pa}_G(i)} = (y^k_i, z^j_i)\ \}|, \\ d_{ij} &= |\{1 \leq \ell \leq d;\ x^\ell_{\mathrm{pa}_G(i)} = z^j_i\ \}|. \end{aligned}$$

Observe that it follows from the definition that $d_{ij} = \sum_{k=1}^{r(i)} d_{ijk}$ for every i, j and that $\sum_{j=1}^{q(i,G)} d_{ij} = d$ for every $i \in N$.

Likelihood function: the sampling distribution has the following density with respect to the arithmetic measure on $\underbrace{\mathsf{X}_N \times \ldots \times \mathsf{X}_N}_{d-\text{times}} \equiv \mathsf{DATA}(N, d)$:

$$\forall\, \boldsymbol{\theta} \in \Theta_G \ \ \forall\, D \in \mathsf{DATA}(N, d) \qquad p_{\boldsymbol{\theta}}(D) = \prod_{i \in N} \prod_{j=1}^{q(i,G)} \prod_{k=1}^{r(i)} \theta_{ijk}^{d_{ijk}} \ \equiv\ L(\boldsymbol{\theta}, D)\,. \tag{16}$$

To verify the formula (16) one "assumes" that D is a random sample from the theoretical distribution $P^{\boldsymbol{\theta}}$. In other words, one considers the d-multiple product of $P^{\boldsymbol{\theta}}$ (with the density $p^{\boldsymbol{\theta}}$) on $\mathsf{X}_N \times \ldots \times \mathsf{X}_N$. Thus, we write:

$$\begin{aligned} p_{\boldsymbol{\theta}}(D) &= \prod_{\ell=1}^{d} p^{\boldsymbol{\theta}}(x^\ell) = \prod_{\ell=1}^{d} \prod_{i \in N} \prod_{j=1}^{q(i,G)} \prod_{k=1}^{r(i)} \theta_{ijk}^{\delta(i,j,k|x^\ell)} = \prod_{i \in N} \prod_{j=1}^{q(i,G)} \prod_{k=1}^{r(i)} \prod_{\ell=1}^{d} \theta_{ijk}^{\delta(i,j,k|x^\ell)} \\ &= \prod_{i \in N} \prod_{j=1}^{q(i,G)} \prod_{k=1}^{r(i)} \theta_{ijk}^{\sum_{\ell=1}^{d} \delta(i,j,k|x^\ell)} = \prod_{i \in N} \prod_{j=1}^{q(i,G)} \prod_{k=1}^{r(i)} \theta_{ijk}^{d_{ijk}}\,. \end{aligned}$$

Here, the second equality can be justified by the same arguments as in (the proof from) § 6.3 and the last equality follows from the definition of d_{ijk}.

Note that it follows from the formula (16) that the class of densities $\{p_{\boldsymbol{\theta}}(D);\ \boldsymbol{\theta} \in \Theta_G\}$ is also an exponential family.[18] The corresponding sufficient statistic for that class of distributions is then the vector statistic

$$\vec{\boldsymbol{d}} = [d_{ijk}]_{\,i \in N,\ j \in \{1,\ldots,q(i,G)\},\ k \in \{1,\ldots,r(i)\}}\,.$$

[18] One can use the similar arguments to those from § 6.3 in the case of the theoretical distribution; just the sample space X_N is replaced by $\mathsf{X}_N \times \ldots \times \mathsf{X}_N$.

In other words, $\vec{\boldsymbol{d}}$ is the collection of marginal counts for all possible parent configurations:

$\forall\, i \in N,\ \forall\, j \in \{1, \ldots, q(i,G)\} \quad \vec{\boldsymbol{d}}_{ij} \equiv [d_{ijk}]_{k=1}^{r(i)}$ is the respective marginal table of counts.

Remark One can also easily derive a formula for the distribution on possible (joint) contingency tables $\boldsymbol{d} : \mathsf{X}_N \to \{0, \ldots, d\}$, $\sum_{x \in \mathsf{X}_N} \boldsymbol{d}(x) = d$.[19] In fact, the "probability" of a particular contingency table $\boldsymbol{d}$ can be obtained as the sum of "probabilities" of (distinct) databases D that lead to $\boldsymbol{d}$. Since all these "probabilities" are the same[20] it is enough to multiply the shared "probability" by the number of these databases, which is a special type of the multinomial coefficient:

$$p_{\boldsymbol{\theta}}(\boldsymbol{d}) = \frac{d!}{\prod_{x \in \mathsf{X}_N} d_{[x]}!} \cdot \prod_{i \in N} \prod_{j=1}^{q(i,G)} \prod_{k=1}^{r(i)} \theta_{ijk}^{d_{ijk}} \ \equiv\ \check{L}(\boldsymbol{\theta}, \boldsymbol{d})\,.$$

7 Bayesian model for a Bayesian network structure

A further step is to enrich the statistical model of a BN structure by a prior probability on the parameter space, that is, to introduce a Bayesian experiment in the sense of Bayesian terminology from § 5.1. The aim of this section is to introduce a whole class of such prior distributions, which appears to be a conjugate family to the considered statistical model.

Throughout this section we keep the conventions from § 6. Thus, we fix an acyclic directed graph G over N and the sample space $\mathsf{X}_N \equiv \prod_{i \in N} \mathsf{X}_i$ with $2 \leq |\mathsf{X}_i| < \infty$ for each $i \in N$. Moreover, we consider the parameter space $\Theta_G = \prod_{i \in N} \prod_{j=1}^{q(i,G)} \Theta_{(ij)}$ of the product form and the system of (theoretical) distributions $\{p^{\boldsymbol{\theta}};\, \boldsymbol{\theta} \in \Theta_G\}$ on X_N given by (15).

7.1 Assumptions

We are going to formulate and comment three assumptions on the prior distribution on Θ_G:

- *parameter independence*,
- *local Dirichlet*, and
- *hyper-consistency*.

1 **Parameter independence assumption** is that the prior probability is a product measure on $\prod_{i \in N} \prod_{j=1}^{q(i,G)} \Theta_{(ij)}$:

$$\pi = \prod_{i \in N} \prod_{j=1}^{q(i,G)} \pi_{(ij)} \quad \text{where } \pi_{(ij)} \text{ is a probability measure on } \Theta_{(ij)}.$$

2 **Local Dirichlet assumption** is that every component $\pi_{(ij)}$ in this product measure is a Dirichlet distribution on the (open) probability simplex $\Theta_{(ij)}$:

$$\forall\, i \in N,\ \forall\, j \in \{1, \ldots, q(i,G)\} \quad \pi_{(ij)} \sim \mathcal{D}([\alpha_{ijk}]_{k=1}^{r(i)}) \text{ for some parameters } \alpha_{ijk} > 0.$$

[19] This is the analogue of the multinomial distribution – c.f. § 3.2.

[20] This follows from the formula (16).

Both these assumptions were already introduced by Spiegelhalter and Lauritzen in [24], where they presented their Bayesian scheme for learning parameters of the statistical model of a discrete BN structure. The first assumption, called *global and local independence* in [24],[21] is quite natural because it allows one to decompose the "global" Bayesian experiment into a system of "local" Bayesian experiments with parameter spaces $\Theta_{(ij)}$. The second assumption is also natural from this point of view because then every "local" statistical model is a "classic" discrete statistical model for a single outcome with parameter space $\Theta_{(ij)}$ and the sample space X_i (see § 3). A conjugate family to this class of distributions is the class of Dirichlet distributions on $\Theta_{(ij)}$ (see § 5.2.1). These assumptions were later taken over by other researchers, including Heckerman, Geiger and Chickering [17]. [22]

To write elegant formulas for the predictive probability and the posterior probabilities the following notational convention is suitable.

Convention III
Under the above-mentioned assumptions we will use the following shorthand notation:

$$\forall\ i \in N,\ \forall\, j \in \{1, \ldots, q(i, G)\} \quad \alpha_{ij}\ = \sum_{k=1}^{r(i)} \alpha_{ijk}.$$

Lemma 4 *If the assumptions* $\boxed{1}$ *and* $\boxed{2}$ *be fulfilled and* $D \in \mathsf{DATA}(N, d)$ *is a database of the length* $d \geq 1$ *then the predictive probability of* D *is given by the formula*

$$p(D) = \prod_{i \in N} \prod_{j=1}^{q(i,G)} \frac{\Gamma(\alpha_{ij})}{\prod_{k=1}^{r(i)} \Gamma(\alpha_{ijk})} \cdot \frac{\prod_{k=1}^{r(i)} \Gamma(\alpha_{ijk} + d_{ijk})}{\Gamma(\alpha_{ij} + d_{ij})}\,, \tag{17}$$

and the posterior probability is again the product of Dirichlet distributions, more specifically

$$\pi(*|D) \sim \prod_{i \in N} \prod_{j=1}^{q(i,G)} \mathcal{D}([\alpha_{ijk} + d_{ijk}]_{k=1}^{r(i)})\,. \tag{18}$$

The proof is, in fact, the modification of arguments from § 5.2.1, where, for every pair (i, j), one considers a localized Bayesian experiment.

Proof. The dominating measure on $\Theta_G \equiv \prod_{i \in N} \prod_{j=1}^{q(i,G)} \Theta_{(ij)}$ is the product of dominating measures $\mu_{\Theta_{(ij)}}$ for local Dirichlet distributions (see § 4.1), while the dominating measure on $\mathsf{DATA}(N, d) \equiv (\mathsf{X}_N)^d$ is the arithmetic measure. Then the joint density (with respect to their product) is, by the assumptions $\boxed{1}$ and $\boxed{2}$, and the formulas (6) and (16), as follows:

$$\begin{aligned}
\Pi(\boldsymbol{\theta}, D) &= \left[\prod_{i \in N} \prod_{j=1}^{q(i,G)} \frac{\Gamma(\sum_{k=1}^{r(i)} \alpha_{ijk})}{\prod_{k=1}^{r(i)} \Gamma(\alpha_{ijk})} \cdot \prod_{k=1}^{r(i)} \theta_{ijk}^{\alpha_{ijk}-1}\right] \cdot \left(\prod_{i \in N} \prod_{j=1}^{q(i,G)} \prod_{k=1}^{r(i)} \theta_{ijk}^{d_{ijk}}\right) \\
&= \prod_{i \in N} \prod_{j=1}^{q(i,G)} \frac{\Gamma(\alpha_{ij})}{\prod_{k=1}^{r(i)} \Gamma(\alpha_{ijk})} \cdot \prod_{k=1}^{r(i)} \theta_{ijk}^{\alpha_{ijk}+d_{ijk}-1}\,.
\end{aligned}$$

[21] More specifically, by *global independence* was meant the assumption that π decomposes with respect to variables $i \in N$ and by *local independence* was meant the assumption that, for each $i \in N$, the respective component further decomposes with respect to the parent configurations.

[22] For example, Assumption 2 in [17] is parameter independence and Assumption 4 in [17] is local Dirichlet.

Now, for fixed D, the predictive probability $p(D)$ is obtained from that by integrating with respect to the corresponding dominating measure:

$$\begin{aligned} p(D) &= \int_{\Theta_G} \Pi(\boldsymbol{\theta}, D)\ d\mu_{\Theta_G}(\boldsymbol{\theta}) = \prod_{i\in N}\prod_{j=1}^{q(i,G)} \int_{\Theta_{(ij)}} \frac{\Gamma(\alpha_{ij})}{\prod_{k=1}^{r(i)}\Gamma(\alpha_{ijk})} \cdot \prod_{k=1}^{r(i)} \theta_{ijk}^{\alpha_{ijk}+d_{ijk}-1}\ d\mu_{\Theta_{(ij)}}(\theta_{ijk}) \\ &= \prod_{i\in N}\prod_{j=1}^{q(i,G)} \frac{\Gamma(\alpha_{ij})}{\prod_{k=1}^{r(i)}\Gamma(\alpha_{ijk})} \cdot \int_{\Theta_{(ij)}} \prod_{k=1}^{r(i)} \theta_{ijk}^{\alpha_{ijk}+d_{ijk}-1}\ d\mu_{\Theta_{(ij)}}(\theta_{ijk}) \\ &= \prod_{i\in N}\prod_{j=1}^{q(i,G)} \frac{\Gamma(\alpha_{ij})}{\prod_{k=1}^{r(i)}\Gamma(\alpha_{ijk})} \cdot \frac{\prod_{k=1}^{r(i)}\Gamma(\alpha_{ijk}+d_{ijk})}{\Gamma(\sum_{k=1}^{r(i)}\alpha_{ijk}+d_{ijk})}, \quad \text{which is another form of (17).} \end{aligned}$$

Here, the second equality follows from the Fubini theorem and the last one from the formula (10) in Corollary 2.

The formula for the posterior density can be obtained by dividing the joint density by the density of the predictive probability:

$$\begin{aligned} \pi(\boldsymbol{\theta}|D) &= \frac{\Pi(\boldsymbol{\theta}, D)}{p(D)} = \frac{\prod_{i\in N}\prod_{j=1}^{q(i,G)} \frac{\Gamma(\alpha_{ij})}{\prod_{k=1}^{r(i)}\Gamma(\alpha_{ijk})} \cdot \prod_{k=1}^{r(i)} \theta_{ijk}^{\alpha_{ijk}+d_{ijk}-1}}{\prod_{i\in N}\prod_{j=1}^{q(i,G)} \frac{\Gamma(\alpha_{ij})}{\prod_{k=1}^{r(i)}\Gamma(\alpha_{ijk})} \cdot \frac{\prod_{k=1}^{r(i)}\Gamma(\alpha_{ijk}+d_{ijk})}{\Gamma(\alpha_{ij}+d_{ij})}} \\ &= \prod_{i\in N}\prod_{j=1}^{q(i,G)} \left[\frac{\Gamma(\alpha_{ij})/\prod_{k=1}^{r(i)}\Gamma(\alpha_{ijk})}{\Gamma(\alpha_{ij})/\prod_{k=1}^{r(i)}\Gamma(\alpha_{ijk})} \cdot \prod_{k=1}^{r(i)} \theta_{ijk}^{\alpha_{ijk}+d_{ijk}-1} \cdot \frac{\Gamma(\alpha_{ij}+d_{ij})}{\prod_{k=1}^{r(i)}\Gamma(\alpha_{ijk}+d_{ijk})} \right] \\ &= \prod_{i\in N}\prod_{j=1}^{q(i,G)} \left\{ \frac{\Gamma(\alpha_{ij}+d_{ij})}{\prod_{k=1}^{r(i)}\Gamma(\alpha_{ijk}+d_{ijk})} \cdot \prod_{k=1}^{r(i)} \theta_{ijk}^{\alpha_{ijk}+d_{ijk}-1} \right\}, \end{aligned}$$

which is, by (6), the density of the product $\prod_{i\in N}\prod_{j=1}^{q(i,G)} \mathcal{D}([\alpha_{ijk}+d_{ijk}]_{k=1}^{r(i)})$. □

The third assumption concerns the parameters of local Dirichlet priors.

3 **Hyperconsistency** is the assumption that the parameters of (local) Dirichlet priors can be obtained by marginalization from a (global) strictly positive function on X_N:

$$\begin{aligned} \exists\, \alpha : \mathsf{X}_N \to (0,\infty) \quad \forall\, i \in N,\ \forall\, j \in \{1,\ldots,q(i,G)\},\ \forall\, k \in \{1,\ldots,r(i)\} \\ \alpha_{ijk} \ = \ \sum \{\, \alpha(x);\ x \in \mathsf{X}_N \ \&\ x_{\{i\}\cup \mathrm{pa}_G(i)} = (y_i^k, z_i^j)\,\}. \end{aligned}$$

Observe the assumption also implies an analogous relation for derived parameters α_{ij}:

$$\forall\, i \in N,\ \forall\, j \in \{1,\ldots,q(i,G)\} \quad \alpha_{ij} = \sum \{\, \alpha(x);\ x \in \mathsf{X}_N \ \&\ x_{\mathrm{pa}_G(i)} = z_i^j\,\}.$$

It is a kind of mutual relation requirent to individual hyper-parameters α_{ijk}. To emphasize that it concerns the hyper-parameters it is named *hyperconsistency*. This terminology was inspired by Dawid and Lauritzen [10], who considered an analogous condition in the context of discrete decomposable (undirected) graphical models.[23] They formulated it in the form of a condition on prior measures for components of their parameter space – see § 3.1 in [10]. There are at least two reasons to accept this assumption:

[23] Every such model can be interpreted as a statistical model of a BN structure.

(i) The previous two assumptions $\boxed{1}$ and $\boxed{2}$ allowed in Lemma 4 to derive formulas for the predictive probability and the posterior density. It follows from those formulas that the corresponding hyper-parameters α_{ijk} can be interpreted as "prior" estimates for the respective contingency tables, say, on the basis of previous measurements.[24] However, the vector of counts $\vec{\boldsymbol{d}} = [d_{ijk}]_{i\in N,\ j\in\{1,\dots,q(i,G)\},\ k\in\{1,\dots,r(i)\}}$ always satisfies the following necessary condition:

$$\begin{aligned}\exists\, \boldsymbol{d} : \mathsf{X}_N \to \{0,\dots,d\} \quad & \forall\, i \in N,\ \forall\, j \in \{1,\dots,q(i,G)\},\ \forall\, k \in \{1,\dots,r(i)\} \\ d_{ijk} \quad &= \quad \sum \{\, \boldsymbol{d}(x);\ x \in \mathsf{X}_N \ \&\ x_{\{i\}\cup pa_G(i)} = (y_i^k, z_i^j)\,\},\end{aligned}$$

where $\boldsymbol{d} : \mathsf{X}_N \to \{0,\dots,d\}$ is the "joint" contingency table. Since the marginal table counts d_{ijk} satisfy this condition and we would like to interpret the numbers α_{ijk} as their prior estimates, it is natural require they satisfy that condition as well.[25]

(ii) The second reason for the acceptance of $\boxed{3}$ is that it is closely related to the requirement *compatibility* of prior distributions for distinct (acyclic directed) graphs. This looks like a natural assumption made to derive Bayesian criteria for learning a BN structure – see later discussion in § 8.2. In fact, the compatibility condition is the requirement that the function α from $\boxed{3}$ is the same for all graphs.

7.2 Hyper-parameter space interpretation

The above-mentioned assumptions allow one to introduce a certain conjugate family of distributions to the statistical model $\{P^{\boldsymbol{\theta}};\ \boldsymbol{\theta} \in \Theta_G\}$ from § 6. There are two ways to introduce the corresponding hyper-parameter space. The one which is chosen here is seemingly not related to the graph G and is more elegant, but it does not lead to a one-to-one correspondence between elements of the space and distributions. Later we relate it to the other way, which is more complicated, but leads to a one-to-one correspondence between elements of the (alternative) space and distributions.

Hyper-parameter space is the set of strictly positive functions on X_N:

$$\Xi \equiv \{\alpha\,;\ \alpha : \mathsf{X}_N \to (0,\infty)\,\}.$$

Prior distributions: We introduce a class of probability distributions on Θ_G:

$$\pi_G^{\alpha} = \prod_{i\in N} \prod_{j=1}^{q(i,G)} \mathcal{D}([\alpha_{ijk}]_{k=1}^{r(i)}) \qquad \text{for any } \alpha \in \Xi,$$

where the hyper-parameters α_{ijk} are obtained from α by "marginalizing":

$$\begin{aligned}&\forall\, i \in N,\ \forall\, j \in \{1,\dots,q(i,G)\},\ \forall\, k \in \{1,\dots,r(i)\} \\ &\qquad \alpha_{ijk} \quad \equiv \quad \sum \{\, \alpha(x);\ x \in \mathsf{X}_N \ \&\ x_{\{i\}\cup pa_G(i)} = (y_i^k, z_i^j)\,\}. \qquad (19)\end{aligned}$$

[24] Actually, the philosophy of the Bayesian approach consists in the **interpretation of the prior information** as the knowledge obtained by gathering information from previous measurements.

[25] Observe, that, by formula (18), the validity of the condition is kept if for "updated hyper-parameters" $\alpha_{ijk} + d_{ijk}$.

Proposition 5 *Given an acyclic directed graph G over N, the class of distributions*

$$\{\, \pi^{\alpha}_{G};\ \alpha \in \Xi \,\}$$

is a conjugate family to the statistical model $\{P^{\boldsymbol{\theta}};\ \boldsymbol{\theta} \in \Theta_G\}$ of a discrete BN network.

More specifically, for every $\alpha \in \Xi$ and $D \in \mathsf{DATA}(N, d)$, $d \geq 1$ the formula for the predictive probability is (17) and the posterior probability is $\pi^{\alpha+\boldsymbol{d}}$, where $\boldsymbol{d} : \mathsf{X}_N \to \mathbb{Z}^+$ is the (joint) contingency table given by D.

Proof. This follows from Lemma 4 and the fact that the hyper-parameters α_{ijk} can be computed from $\alpha : \mathsf{X}_N \to (0, \infty)$ by the same formula (19) as d_{ijk} from $\boldsymbol{d} : \mathsf{X}_N \to \mathbb{Z}^+$. □

However, the mapping $\alpha \mapsto \pi^{\alpha}_{G}$ is not injective. One can easily observe this:

Lemma 6 *Given $\alpha, \beta \in \Xi$ one has $\pi^{\alpha}_{G} = \pi^{\beta}_{G}$ iff α and β have the same marginals for the (maximal) sets from the system $\{\, \{i\} \cup \mathrm{pa}_G(i)\,;\ i \in N\}$, that is:*

$$\forall\, i \in N \quad \forall\, y \in \mathsf{X}_{\{i\}\cup \mathrm{pa}_G(i)}$$
$$\sum \{\, \alpha(x);\ x \in \mathsf{X}_N \ \&\ x_{\{i\}\cup \mathrm{pa}_G(i)} = y \,\} \quad = \quad \sum \{\, \beta(x);\ x \in \mathsf{X}_N \ \&\ x_{\{i\}\cup \mathrm{pa}_G(i)} = y \,\}\,.$$

Proof. First, realize that the condition above can be re-formulated in this form: $\alpha_{ijk} = \beta_{ijk}$ for any $i \in N$, $j \in \{1, \ldots, q(i, G)\}$ and $k \in \{1, \ldots, r(i)\}$. This is because, given $i \in N$, any pair (j, k) encodes a parent-node configuration $(y^k_i, z^j_i) \in \mathsf{X}_{\{i\}\cup \mathrm{pa}_G(i)}$.

Moreover, $\pi^{\alpha}_{G} = \pi^{\beta}_{G}$ iff, for any i and j, one has $\mathcal{D}([\alpha_{ijk}]^{r(i)}_{k=1}) = \mathcal{D}([\beta_{ijk}]^{r(i)}_{k=1})$. Thus, the sufficiency of the condition is evident and the necessity follows from the one-to-one correspondence between parameters and distributions for the class of Dirichlet distributions – see the remark following Corollary 2 (on page 15). □

Thus, to have a one-to-one correspondence between the elements of the hyper-parameter space and priors one should introduce the space as follows. Let $\mathcal{A}(G)$ denote the system of *maximal sets* (with respect to inclusion) in the class $\{\, \{i\} \cup \mathrm{pa}_G(i)\,;\ i \in N\}$. Then put:

$$\Xi_G \quad = \quad \Big\{\, [\alpha_A]_{A\in\mathcal{A}(G)}\,;\ \text{where } \alpha_A : \mathsf{X}_A \to (0,\infty) \text{ are such that } \exists\, \alpha : \mathsf{X}_N \to (0,\infty)$$
$$\text{with } \alpha_A(y) = \sum \{\, \alpha(x);\ x \in \mathsf{X}_N \ \&\ x_{\{i\}\cup \mathrm{pa}_G(i)} = y \,\} \text{ for } y \in \mathsf{X}_A, A \in \mathcal{A}(G) \Big\}\,.$$

Given $\boldsymbol{\alpha} \equiv [\alpha_A]_{A\in\mathcal{A}(G)} \in \Xi_G$ the corresponding hyper-parameters are computed by:

$$\forall\, i \in N,\ \forall\, j \in \{1, \ldots, q(i, G)\},\ \forall\, k \in \{1, \ldots, r(i)\}$$
$$\alpha_{ijk} \quad \equiv \quad \sum \{\, \alpha_A(x);\ x \in \mathsf{X}_A \ \&\ x_{\{i\}\cup \mathrm{pa}_G(i)} = (y^k_i, z^j_i) \,\}$$
$$\text{for some } A \in \mathcal{A}(G) \text{ such that } \{i\} \cup pa_G(i) \subseteq A.$$

Note that the assumption $\boldsymbol{\alpha} \in \Xi_G$ implies that the values of α_{ijk} do not depend on the choice of $A \in \mathcal{A}(G)$ satisfying $\{i\} \cup pa_G(i) \subseteq A$. The respective class of (prior) distributions on Θ_G is defined analogously:

$$\pi^{\boldsymbol{\alpha}}_{G} = \prod_{i\in N} \prod_{j=1}^{q(i,G)} \mathcal{D}([\alpha_{ijk}]^{r(i)}_{k=1}) \qquad \text{for any } \boldsymbol{\alpha} \in \Xi_G\,.$$

It follows from the construction that

$$\{\,\pi_G^\alpha;\ \alpha\in\Xi\}\ \equiv\ \{\,\pi_G^{\boldsymbol{\alpha}};\ \boldsymbol{\alpha}\in\Xi_G\},,$$

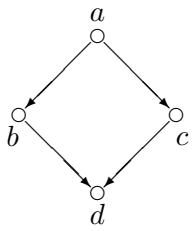
which means Ξ_G can also be viewed as the hyper-parameter space for the conjugate family to the statistical model $\{P^{\boldsymbol{\theta}};\ \boldsymbol{\theta}\in\Theta_G\}$.

Remark (*warning*) The reader may have the temptation to believe that one can perhaps introduce a suitable hyper-parameter space for $\{P^{\boldsymbol{\theta}};\ \boldsymbol{\theta}\in\Theta_G\}$ as follows:

$$\Xi_G^*=\{\alpha^*:\mathsf{X}_N\to(0,\infty)\,;\ \alpha^*\text{ is a (positive) multiple of a distribution in }\mathcal{M}_+(G,\mathsf{X}_N)\,\}.$$

Indeed, this way would ensure, by Lemma 6, that the mapping $\alpha\mapsto p^\alpha$, $\alpha\in\Xi_G^*$ is an injective mapping. On the other hand, the restriction to the class $\{p^\alpha;\ \alpha\in\Xi_G^*\}$ is not a good idea, because it is **not** a conjugate family to $\{P^{\boldsymbol{\theta}};\ \boldsymbol{\theta}\in\Theta_G\}$!

This is an example. Let us put $N=\{a,b,c,d\}$, $\mathsf{X}_i=\{\mathsf{0},\mathsf{1}\}$ for every $i\in N$ and consider the following graph G over N:



The basic idea is to construct a special a probability density $q:\mathsf{X}_N\to(0,1)$, namely the product $q=q_a\times q_{bc}\times q_d$ where q_{bc} is **not** a product on $\mathsf{X}_b\times\mathsf{X}_c$. Then consider any $\alpha^\dagger\in\Xi$ of the form $\alpha^\dagger=k\cdot q$, $k>0$. The argument is that there is no $\beta:\mathsf{X}_N\to(0,\infty)$ which is a multiple of a Markovian distribution from $\mathcal{M}_+(G,\mathsf{X}_N)$ and satisfies $\pi_G^{\alpha^\dagger}=\pi_G^\beta$. Indeed, by Lemma 6, one should have then $\alpha^\dagger_{ab}=\beta_{ab}$, $\alpha^\dagger_{ac}=\beta_{ac}$ and $\alpha^\dagger_{bcd}=\beta_{bcd}$. In particular, for the hypothetical $p\in\mathcal{M}_+(G,\mathsf{X}_N)$ with $\beta=k\cdot p$, one must have $p_{ab}=q_{ab}$, $p_{ac}=q_{ac}$ and $p_{bcd}=q_{bcd}$. Since $q_{ab}=q_a\times q_b$ and $q_{ac}=q_a\times q_c$ one has $a\perp\!\!\!\perp b\,|\,\emptyset\ [p]$ and $a\perp\!\!\!\perp c\,|\,\emptyset\ [p]$. The Markov condition for p (with respect to G) implies $b\perp\!\!\!\perp c\,|\,a\ [p]$. Thus, by basic properties of conditional independence, $p_{abc}=p_a\times p_b\times p_c$, which implies p_{bc} **is** a product on $\mathsf{X}_b\times\mathsf{X}_c$. Then $p_{bc}=q_{bc}$ contradicts the original choice of q.

To show that $\{p^\alpha;\ \alpha\in\Xi_G^*\}$ is not a conjugate family to $\{P^{\boldsymbol{\theta}};\ \boldsymbol{\theta}\in\Theta_G\}$ it suffices to put $\alpha\equiv 1$ (which belongs to Ξ_G^*) and to construct $D\in\mathsf{DATA}(N,d)$ such that, for the corresponding contingency table $\boldsymbol{d}:\mathsf{X}_N\to\mathbb{Z}^+$, the sum $\alpha+\boldsymbol{d}$ has the form $\alpha^\dagger$ mentioned above. In particular, the posterior probability $\pi^{\alpha+\boldsymbol{d}}$ does not belong to $\{p^\alpha;\ \alpha\in\Xi_G^*\}$.

Remark (*conjecture*) This is a natural conjecture: for (independence) equivalent graphs G and H over N and $\alpha\in\Xi$, the prior distribution π_G^α induces by the mapping $\boldsymbol{\theta}\mapsto p^{\boldsymbol{\theta}}$, $\boldsymbol{\theta}\in\Theta_G$ (see Lemma 3 in § 6.2) on $\mathcal{M}_+(G,\mathsf{X}_N)=\mathcal{M}_+(H,\mathsf{X}_N)$ the same distribution as π_H^α by its respective mapping from Θ_H to $\mathcal{M}_+(H,\mathsf{X}_N)$.[26]

[26]This is (maybe) a hint for a possible proof. Owing to transformational characterization of (independence) equivalence one can consider without loss of generality that $G,H\in\mathsf{DAGS}(N)$ differing by legal arrow reversal $a\to b$. Then observe that the transformation $\boldsymbol{\theta}\mapsto\boldsymbol{\vartheta}$ from Θ_G to Θ_H given by $p^{\boldsymbol{\theta}}=p^{\boldsymbol{\vartheta}}$ is "local" for the pair (a,b): it only concerns $\prod_{i\in\{a,b\}}\prod_j\Theta_{(ij)}$. Finally, for given $\alpha:\mathsf{X}_N\to(0,\infty)$, observe that the "local distribution" $\prod_{j_a}\mathcal{D}([\alpha^G_{ajk}])\times\prod_{j_b}\mathcal{D}([\alpha^G_{bjk}])$ is transformed by that local transform to $\prod_{j_b}\mathcal{D}([\alpha^H_{bjk}])\times\prod_{j_a}\mathcal{D}([\alpha^H_{ajk}])$. This is perhaps relevant to Lemma 7.2 from [10].

Remark If one considers, instead the sample space $(\mathsf{X}_N)^d \equiv \mathsf{DATA}(N,d)$, the collection of all possible joint contingency tables $\boldsymbol{d} : \mathsf{X}_N \to \{0,\ldots,d\}$, $\sum_{x\in\mathsf{X}_N} \boldsymbol{d}(x) = d$, then the formula for the predictive probability differs from (17) only by the multinomial coefficient:

$$p^{\alpha}(\boldsymbol{d}) = \frac{d!}{\prod_{x\in\mathsf{X}_N} d_{[x]}!} \cdot \prod_{i\in N} \prod_{j=1}^{q(i,G)} \frac{\Gamma(\alpha_{ij})}{\prod_{k=1}^{r(i)} \Gamma(\alpha_{ijk})} \cdot \frac{\prod_{k=1}^{r(i)} \Gamma(\alpha_{ijk}+d_{ijk})}{\Gamma(\alpha_{ij}+d_{ij})} \,.$$

8 Bayesian criteria for learning a Bayesian network structure

Bayesian approach leads to a class of quality criteria for learning a BN structure. In this section, we first recapitulate what is meant by a quality criterion (for learning a BN structure) and then formally introduce a class of Bayesian criteria defined as the *logarithms of predictive probabilities*.[27] Throughout the section we accept the following notational convention:

Convention IV
Given a non-empty finite set of variables N, the symbol $\mathsf{DAGS}(N)$ will denote the collection of all acyclic directed graphs over N (= having N as the set of nodes).

8.1 The concept of a quality criterion

The method for learning a BN structure by maximizing a quality criterion is based on the following concept:

Definition 5 (score equivalent quality criterion)
By a *quality criterion* for learning a BN structure is meant a real function of two variables, namely of an acyclic directed graph and a database:

$$\mathcal{Q} : \mathsf{DAGS}(N) \times \mathsf{DATA}(N,d) \to \mathbb{R}\,,\ d \geq 1\,.$$

A quality criterion will be named *score equivalent* if it does not distinguish between graphs defining the same statistical model (= the same BN structure):

$$\mathcal{Q}(G,D) \;=\; \mathcal{Q}(H,D) \quad \text{whenever } G,H \in \mathsf{DAGS}(N) \text{ are independence equivalent.}$$

Here, graphs are *independence equivalent* if they induce the same conditional independence structure, that is, define the same collection of conditional independence restrictions through the correspoding separation criterion, described for example in § 3.2.2 of [19].[28]

The interpretation of the value $\mathcal{Q}(G,D)$ for $G \in \mathsf{DAGS}(N)$ and $D \in \mathsf{DATA}(N,d)$, $d \geq 1$ is that it should evaluate how the statistical model determined by G fits the database D. Whether a particular criterion $\mathcal{Q}$ really meets such an intuitive requirement is a more subtle question, closely related to the question of *statistical consistency* of $\mathcal{Q}$ – see [22] or [7] for

[27] These criteria are equivalent to what some other authors call either "*marginal likelihood*" or "marginal probability of data" [9].

[28] Note one can show that if $|\mathsf{X}_i| \geq 2$ for any $i \in N$, which is our assumption from § 6.1, then *Markov equivalence* of acyclic directed graphs (over N) coincides with their independence equivalence – c.f. § 6.1.1 in [27]. More specifically, $\mathcal{M}_+(G,\mathsf{X}_N) = \mathcal{M}_+(H,\mathsf{X}_N)$ for $G,H \in \mathsf{DAGS}(N)$, which means they define the same statistical model in the sense of § 6, iff G and H are independence equivalent.

(informal) definitions of this concept. The topic of statistical consistency of Bayesian quality criteria is omitted in this report; the intention is to deal with it in detail later.

Thus, the learning procedure should consist in maximizing the function $G \mapsto \mathcal{Q}(G, D)$, where $D \in \mathsf{DATA}(N, d)$, $d \geq 1$ is the observed database. The requirement that $\mathcal{Q}$ is score equivalent is then quite natural from the point of view of the purpose of the learning procedure: the aim is learn a BN structure! The term "score equivalent" was taken over from [5]. Note that there are simple graphical characterizations of independence equivalence of acyclic directed graphs – see § 3.2 in [27] for an overview.

The second important technical requirement on quality criteria was formulated in connection with machine learning approach to the maximization problem. Since direct maximization of the function $G \mapsto \mathcal{Q}(G, D)$ seems, at least at first sight, to be infeasible, various *local search methods* have been proposed instead. These methods are applicable to criteria which satisfy the following condition:

Definition 6 (decomposable quality criterion)
A quality criterion $\mathcal{Q}$ will be called (*additively*) *decomposable* if there is 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 $d \geq 1$, such that

$$\mathcal{Q}(G, D) = \sum_{i \in N} q_{i|pa_G(i)}(D_{\{i\} \cup pa_G(i)}) \quad \text{for every } G \in \mathsf{DAGS}(N),\ D \in \mathsf{DATA}(N, d). \tag{20}$$

Here, the symbol D_A for $\emptyset \neq A \subseteq N$ and $D \in \mathsf{DATA}(N, d)$ denotes the a *projection* of the database $D : x^1, \ldots, x^d$ onto X_A, that is, the sequence of respective marginal configurations $D_A : x^1_A, \ldots, x^d_A$.
The criterion $\mathcal{Q}$ will be called *strongly decomposable*, if, moreover, the functions $q_{i|B}$ only depend on the marginal table of counts $\boldsymbol{d}_{\{i\} \cup pa_G(i)}$ given by $D_{\{i\} \cup pa_G(i)}$.[29]

This definition was basically taken over from § 2.3 of [7]. However, in that paper, Chickering was not completely clear what he means by "data" in his definition. His word description indicates that he has probably in mind data represented in the form of a sample (see item **B** in § 2.1). On the other hand, later (in § 4.1 of [7]), he restricts his attention solely to the criteria that do not depend on the order of items in a database; this more corresponds to "data" in the form of a contingency table (see item **C** in § 2.1). In this context, the reader has natural implicit tendency to interpret Chickering's definition in the sense that the components $q_{i|pa_G(i)}$ in (20) also should not depend on the order of items in the database. Since this requirement is indeed a stronger condition, but valid for most criteria used in practice, I have distinguished that stronger condition terminologically.[30]

Another potential source of misunderstanding in connection with the concept of a decomposable criterion is that some other authors [16, 9] consider "multiplicative" versions of (decomposable) quality criteria. These are criteria $\tilde{\mathcal{Q}} : \mathsf{DAGS}(N) \times \mathsf{DATA}(N, d) \to (0, \infty)$, $d \geq 1$ that *factorize* relative to the graph, which means

$$\tilde{\mathcal{Q}}(G, D) = \prod_{i \in N} \tilde{q}_{i|pa_G(i)}(D_{\{i\} \cup pa_G(i)}) \quad \text{for } G \in \mathsf{DAGS}(N),\ D \in \mathsf{DATA}(N, d).$$[31]

[29] This essentially means the functions $q_{i|B}$ do not depend on the order of items in the database.
[30] One can show using the results of § 8.4.2 in [27] that if a criterion is decomposable and does not depend on the order of items in a database then it is strongly decomposable.
[31] That means, the criterion decomposes *multiplicatively*, not additively.

However, this is just a small technical difference, because on can switch between the additive and multiplicative version of a decomposable criterion easily: $\mathcal{Q}(G,D) \equiv \ln \tilde{\mathcal{Q}}(G,D)$. Since the logarithmic transformation is order-preserving the task to maximize $G \mapsto \tilde{\mathcal{Q}}(G,D)$ is equivalent to the task to maximize $G \mapsto \mathcal{Q}(G,D)$.

The following auxiliary observation is useful – for the proof see Lemma 8.3 in [27].

Lemma 7 *A quality criterion $\mathcal{Q}$ is score equivalent and strongly decomposable iff there exists a collection of real functions $\{\, \boldsymbol{t}_A;\, A \subseteq N\}$, each $\boldsymbol{t}_A$ depending on the respective marginal contingency table $\boldsymbol{d}_A : \mathsf{X}_A \to \{0,1,\ldots,d\}$, $d \geq 1$ computed from D_A, such that*

$$\mathcal{Q}(G,D) \;=\; \sum_{i \in N} \;\big\{\; \boldsymbol{t}_{\{i\}\cup \mathrm{pa}_G(i)}(\boldsymbol{d}_{\{i\}\cup \mathrm{pa}_G(i)}) \;-\; \boldsymbol{t}_{\mathrm{pa}_G(i)}(\boldsymbol{d}_{\mathrm{pa}_G(i)}) \;\big\} \tag{21}$$

for every $G \in \mathsf{DAGS}(N)$ and $D \in \mathsf{DATA}(N,d)$.[32]

8.2 Compatibility assumption

The predictive probability depends on the (choice of the) prior distribution, which is, however, defined on a specific parameter space Θ_G (of the statistical model) given by an acyclic directed graph G. However, since we are in the situation we are supposed to compare different structural models (= corresponding to different graphs), we have to accept some assumption of *compatibility* of prior distributions for distinct structural models. In our context, it is the following assumption:

> $\exists\, \alpha : \mathsf{X}_N \to (0,\infty)$ such that, for every $G \in \mathsf{DAGS}(N)$, the prior distribution π^α_G on the parameter space Θ_G is specified as follows:
>
> - the hyper-parameters of priors are given by (marginalization from) α:
>
> $$\forall\, i \in N,\ \forall\, j \in \{1,\ldots,q(i,G)\},\ \forall\, k \in \{1,\ldots,r(i)\}$$
> $$\alpha^G_{ijk} \;\equiv\; \sum \{\, \alpha(x);\ x \in \mathsf{X}_N \ \&\ x_{\{i\}\cup \mathrm{pa}_G(i)} = (y^k_i, z^j_i)\,\},$$
>
> - and the prior is the product of corresponding Dirichlet distributions:
>
> $$\pi^\alpha_G = \prod_{i \in N} \prod_{j=1}^{q(i,G)} \mathcal{D}([\alpha^G_{ijk}]_{k=1}^{r(i)})\,.$$

In particular, for every $G \in \mathsf{DAGS}(N)$, the assumptions 1-3 from § 7 are fulfilled. The function $\alpha : \mathsf{X}_N \to (0,\infty)$ will be called the (shared) *hyper-potential*.

Having fixed a hyper-potential $\alpha : \mathsf{X}_N \to (0,\infty)$ the corresponding (Bayesian) quality criterion, considered in this report, is given by

$$\mathsf{LML}_\alpha(G,D) \;=\; \ln\, p^\alpha_G(D) \qquad \text{for } G \in \mathsf{DAGS}(N),\ D \in \mathsf{DATA}(N,d), \tag{22}$$

where $p^\alpha_G(D)$ denotes the predictive probability of D under the assumption the prior distribution on Θ_G is π^α_G.[33] For the reasons that become evident later in § 8.3, we consider

[32] The marginal table of counts $\boldsymbol{d}_\emptyset$ for $A = \emptyset$ is the function on one-element set $\mathsf{X}_\emptyset$ having the value d. Thus, the value $\boldsymbol{t}_\emptyset(\boldsymbol{d}_\emptyset)$ is, in fact, a constant (= it only depends on d and the fixed joint sample space).

[33] The abbreviation LML stands for "*logarithm of the marginal likehood*".

an additive version of the criterion. Other authors [17] consider the multiplicative version instead, which is just the predictive probability $p^{\alpha}_{G}(D)$.

The question of the choice of prior distributions on spaces Θ_G, $G \in \mathsf{DAGS}(N)$ was discussed in detail by Heckerman, Geiger and Chickering [17]. The main result of that paper is that the above-mentioned compatibility condition follows naturally from considerations about what are reasonable requirements on (Bayesian) quality criteria. Those requirements are formulated in [17] in the form of assumptions and the compatibility condition is derived as their consequence. They also give a formula for the respective (multiplicative version of the) quality criterion, called "*BDe-metric*". Moreover, the compatibility condition is formulated in [17] in slightly different form. More specifically, they require $\alpha = \alpha_+ \cdot p$, where $p : \mathsf{X}_N \to (0,1)$ is a probability density and $\alpha_+ > 0$. This formulation leads to the following interpretation. The number α_+ is called "user's equivalent sample size" and the density p is named "prior network".

The compatibility issue was also dealt with by Dawid and Lauritzen in § 6.2 of [10], in the context of (undirected) decomposable graphical models. They basically suggest the same procedure for the choice of prior distributions for different structural models.

The basic observation concerning the above-mentioned criterion is as follows.

Proposition 8 *If the compatibility assumption is valid and $\alpha : \mathsf{X}_N \to (0,\infty)$ is the given hyper-potential then the corresponding Bayesian criterion is given by the following formula: for every $G \in \mathsf{DAGS}(N)$ and $D \in \mathsf{DATA}(N,d)$*

$$\mathsf{LML}_{\alpha}(G,D) \;=\; \sum_{i\in N}\sum_{j=1}^{q(i,G)} \left\{ \ln \frac{\Gamma(\alpha^{G}_{ij})}{\Gamma(\alpha^{G}_{ij}+d_{ij})} \;-\; \sum_{r=1}^{r(i)} \ln \frac{\Gamma(\alpha^{G}_{ijk})}{\Gamma(\alpha^{G}_{ijk}+d_{ijk})} \right\}, \tag{23}$$

where α^{G}_{ijk} are corresponding hyper-parameters and d_{ijk} the corresponding marginal counts. Moreover, the criterion is score equivalent and strongly decomposable. More specifically, if we put, for $A \subseteq N$,

$$\boldsymbol{t}^{\alpha}_{A}(\boldsymbol{d}_A) \;=\; \sum_{y\in \mathsf{X}_A} \ln \frac{\Gamma(\alpha_A(y)+\boldsymbol{d}_A(y))}{\Gamma(\alpha_A(y))}\,, \tag{24}$$

where $\alpha_A : \mathsf{X}_A \to (0,\infty)$ denotes the marginal hyper-potential and $\boldsymbol{d}_A : \mathsf{X}_A \to \{0,1,\ldots,d\}$ the marginal contingency table, then one has

$$\mathsf{LML}_{\alpha}(G,D) \;=\; \sum_{i\in N} \left\{\; \boldsymbol{t}^{\alpha}_{\{i\}\cup pa_G(i)}(\boldsymbol{d}_{\{i\}\cup pa_G(i)}) \;-\; \boldsymbol{t}^{\alpha}_{pa_G(i)}(\boldsymbol{d}_{pa_G(i)}) \;\right\} \tag{25}$$

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

Proof. The formula for the predictive probability $p^{\alpha}_{G}(D)$ is given by (17) in Lemma 4. By taking its logarithm one gets

$$\ln\, p^{\alpha}_{G}(D) \;=\; \sum_{i\in N}\sum_{j=1}^{q(i,G)} \ln \left(\frac{\Gamma(\alpha^{G}_{ij})}{\prod_{k=1}^{r(i)}\Gamma(\alpha^{G}_{ijk})} \cdot \frac{\prod_{k=1}^{r(i)}\Gamma(\alpha^{G}_{ijk}+d_{ijk})}{\Gamma(\alpha^{G}_{ij}+d_{ij})} \right),$$

and then, by standard re-writing, the formula (23) is obtained. Morever, it can written as follows (the superscript G with α_{ijk}'s is dropped):

$$\begin{aligned}
\mathsf{LML}_\alpha(G,D) &= \sum_{i\in N}\sum_{j=1}^{q(i,G)}\left\{\sum_{k=1}^{r(i)}\ln\frac{\Gamma(\alpha_{ijk}+d_{ijk})}{\Gamma(\alpha_{ijk})} - \ln\frac{\Gamma(\alpha_{ij}+d_{ij})}{\Gamma(\alpha_{ij})}\right\}\\
&= \sum_{i\in N}\left\{\sum_{j=1}^{q(i,G)}\sum_{k=1}^{r(i)}\ln\frac{\Gamma(\alpha_{ijk}+d_{ijk})}{\Gamma(\alpha_{ijk})} - \sum_{j=1}^{q(i,G)}\ln\frac{\Gamma(\alpha_{ij}+d_{ij})}{\Gamma(\alpha_{ij})}\right\}\\
&= \sum_{i\in N}\left\{\sum_{y\in \mathsf{X}_{\{i\}\cup pa_G(i)}}\ln\frac{\Gamma(\,\alpha_{\{i\}\cup pa_G(i)}(y)+\boldsymbol{d}_{\{i\}\cup pa_G(i)}(y)\,)}{\Gamma(\,\alpha_{\{i\}\cup pa_G(i)}(y)\,)}\right.\\
&\qquad\left. - \sum_{y\in \mathsf{X}_{pa_G(i)}}\ln\frac{\Gamma(\,\alpha_{pa_G(i)}(y)+\boldsymbol{d}_{pa_G(i)}(y)\,)}{\Gamma(\,\alpha_{pa_G(i)}(y)\,)}\right\}.
\end{aligned}$$

Here, the third equality is valid because, for fixed $i\in N$, pairs (j,k) encode elements of the corresponding marginal sample space $\mathsf{X}_{\{i\}\cup pa_G(i)}$, while j's encode elements of $\mathsf{X}_{pa_G(i)}$. By (24), the formula (25) is obtained. The fact that LML_α is score equivalent and strongly decomposable then follows from Lemma 7. □

Remark If one considers, alternatively, the predictive probability on the collection of joint contingency tables, then the corresponding formula has one more term, namely the logarithm of the multinomial coefficient (c.f. the remark concluding § 7):

$$\ln p^\alpha(\boldsymbol{d}) = \ln\frac{d!}{\prod_{x\in\mathsf{X}_N} d_{[x]}!} + \underbrace{\sum_{i\in N}\sum_{j=1}^{q(i,G)}\left\{\ln\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+d_{ij})} - \sum_{r=1}^{r(i)}\ln\frac{\Gamma(\alpha_{ijk})}{\Gamma(\alpha_{ijk}+d_{ijk})}\right\}}_{\mathsf{LML}_\alpha(G,D)}.$$

This modification, however, is not suitable, because the additional term "destroys" strong decomposability of the criterion, which is quite desirable property. Indeed, if $\ln p^\alpha(\boldsymbol{d})$ had been strongly decomposable then, because LML_α is strongly decomposable, their difference (which, in fact, does not depend on G at all)

$$\breve{\mathcal{Q}}(G,D) = \ln\frac{d!}{\prod_{x\in\mathsf{X}_N} d_{[x]}!} = \ln(d!) - \sum_{x\in\mathsf{X}_N}\ln(d_{[x]}!)$$

would have been strongly decomposable.

This is a counter-example: put $N=\{a,b\}$, $\mathsf{X}_i=\{0,1\}$ for every $i\in N$ and consider the empty graph G over N. Decomposability hypothesis for $\breve{\mathcal{Q}}$ leads to the requirement there exist functions q_a and q_b (of the respective marginal contingency tables) that $\breve{\mathcal{Q}}(G,D) = q_a(\boldsymbol{d}_{\{a\}}) + q_b(\boldsymbol{d}_{\{b\}})$. Let us consider two databases D^1 and D^2 of the length $d=4$ with the following joint contingency tables:

$$\boldsymbol{d}^1:\ \begin{array}{ll}(0,0),(1,1) & \mapsto\ 2\\ (0,1),(1,0) & \mapsto\ 0\end{array}\qquad \boldsymbol{d}^2:\ (0,0),(1,1),(0,1),(1,0)\ \mapsto\ 1\ .$$

Clearly, $\boldsymbol{d}^1_{\{a\}} = \boldsymbol{d}^2_{\{a\}}$ and $\boldsymbol{d}^1_{\{b\}} = \boldsymbol{d}^2_{\{b\}}$. Thus, if $\breve{\mathcal{Q}}$ had been strongly decomposable then it would have been $\breve{\mathcal{Q}}(G, D^1) = \breve{\mathcal{Q}}(G, D^2)$. But direct computation gives

$$\begin{aligned}\breve{\mathcal{Q}}(G, D^1) &= \ln(4!) - 2\ln(2!) - 2\ln(0!) = 3\ln 2 + \ln 3 - 2\ln 2 - 2\ln 1 = \ln 2 + \ln 3,\\ \breve{\mathcal{Q}}(G, D^2) &= \ln(4!) - 4\ln(1!) = 3\ln 2 + \ln 3 - 4\ln 1 = 3\ln 2 + \ln 3.\end{aligned}$$

Hence, $\breve{\mathcal{Q}}(G, D^1) \neq \breve{\mathcal{Q}}(G, D^2)$, which is a contradiction.

8.3 Formula for the data vector

The basic idea of an algebraic approach to learning a BN structure presented in Chapter 8 of [27] is to represent both the BN structure and the database by a special real vector. The algebraic representative of the BN structure is a certain integral (= integer-valued) vector. These integral vectors are called imsets in [27]:

Definition 7 (standard imset)
An *imset* u over a (non-empty finite) set of variables N is a function $u : \mathcal{P}(N) \mapsto \mathbb{Z}$, where $\mathcal{P}(N) \equiv \{A; A \subseteq N\}$ denotes the power set of N. Given $A \subseteq N$, the symbol δ_A will denote a special imset given by:

$$\delta_A(B) = \begin{cases} 1 & \text{if } B = A, \\ 0 & \text{if } B \neq A, \end{cases} \qquad \text{for } \ B \subseteq N.$$

Then, the *standard imset* for an acyclic directed graph $G \in \mathsf{DAGS}(N)$ is given by the formula

$$u_G = \delta_N - \delta_\emptyset + \sum_{i \in N} \left\{ \delta_{pa_G(i)} - \delta_{\{i\} \cup pa_G(i)} \right\}. \tag{26}$$

We will regard every imset as a vector whose components are integers and are indexed by subsets of N. Note the standard imset is uniquely determined representative of the BN structure: Corollary 7.1 in [27] says that, given $G, H \in \mathsf{DAGS}(N)$, one has $u_G = u_H$ iff they are independence equivalent.

Actually, any real function $m : \mathcal{P}(N) \to \mathbb{R}$ can be interpreted as a (real) vector in the same way. The scalar product of two vectors of this type will be denoted as follows:

$$\langle m, u \rangle \equiv \sum_{A \subseteq N} m(A) \cdot u(A).$$

The point is that every database $D \in \mathsf{DATA}(N, d)$, $d \geq 1$ can be represented by a real vector of this type. The following observation is a simple consequence of Lemma 7:

Corollary 9 *Let $\mathcal{Q}$ be a score equivalent and strongly decomposable criterion. Then there exist a real function $s^{\mathcal{Q}} : \mathsf{DATA}(N, d) \to \mathbb{R}$ and a vector function $t^{\mathcal{Q}} : \mathsf{DATA}(N, d) \to \mathbb{R}^{\mathcal{P}(N)}$ such that*

$$\mathcal{Q}(G, D) = s^{\mathcal{Q}}_D - \langle t^{\mathcal{Q}}_D, u_G \rangle \quad \textit{for every } G \in \mathsf{DAGS}(N),\ D \in \mathsf{DATA}(N, d). \tag{27}$$

Here, $s^{\mathcal{Q}}_D$ and $t^{\mathcal{Q}}_D$ denote the respective values of the above functions for $D \in \mathsf{DATA}(N, d)$ and, for every $A \subseteq N$, $t^{\mathcal{Q}}_D(A)$ only depends on the marginal contingency table $\boldsymbol{d}_A$.

Proof. Let us consider the formula (21), where we put $t^{\mathcal{Q}}_D(A) \equiv \boldsymbol{t}_A(\boldsymbol{d}_A)$ for any $A \subseteq N$:

$$\begin{aligned}
\mathcal{Q}(G,D) &= \sum_{i\in N} \{\langle t^{\mathcal{Q}}_D\,,\,\delta_{\{i\}\cup pa_G(i)}\rangle - \langle t^{\mathcal{Q}}_D\,,\,\delta_{pa_G(i)}\rangle\} \;=\; \langle t^{\mathcal{Q}}_D\,,\,\sum_{i\in N}\{\delta_{\{i\}\cup pa_G(i)} - \delta_{pa_G(i)}\}\,\rangle \\
&= \langle t^{\mathcal{Q}}_D\,,\,\delta_N - \delta_\emptyset \;-\; \underbrace{[\,\delta_N - \delta_\emptyset + \sum_{i\in N}\{\delta_{pa_G(i)} - \delta_{\{i\}\cup pa_G(i)}\}\,]}_{u_G}\rangle \\
&= \underbrace{\langle t^{\mathcal{Q}}_D\,,\,\delta_N - \delta_\emptyset\rangle}_{s^{\mathcal{Q}}_D} - \langle t^{\mathcal{Q}}_D\,,\,u_G\rangle\,.
\end{aligned}$$

This gives the required formula and $t^{\mathcal{Q}}_D(A)$ only depends on $\boldsymbol{d}_A$. □

The obtained formula (27) basically says the criterion can be viewed as an affine function (= a linear function plus a constant) of the standard imset.

Definition 8 (data vector)
Given a score equivalent and strongly decomposable criterion $\mathcal{Q}$ and $D \in \mathsf{DATA}(N,d)$, $d \geq 1$ by the *data vector* relative to $\mathcal{Q}$ will be named any vector $t^{\mathcal{Q}}_D$ that satisfies (27). The function $s^{\mathcal{Q}}_D$ from (27) will be called the *saturating function* of $\mathcal{Q}$.

Note that the saturating function is uniquely determined by the criterion: it is the value of the criterion for (any) complete acyclic directed graph over N. On the other hand, the data vector is not uniquely determined but it becomes unique if one requires the following standardization condition:

$$t^{\mathcal{Q}}_D(A) = 0 \qquad \text{for any } A \subseteq N \text{ with } |A| \leq 1.$$

For the corresponding arguments see Lemma 8.7 in [27].

Now, Proposition 8 allows one to derive elegant formulas for the data vector:

Corollary 10 *If the compatibility assumption is valid and $\alpha : \mathsf{X}_N \to (0,\infty)$ is the given hyper-potential then the formula*

$$\check{t}^{\mathsf{LML},\alpha}_D(A) = \sum_{y\in\mathsf{X}_A} \ln \frac{\Gamma(\alpha_A(y) + \boldsymbol{d}_A(y))}{\Gamma(\alpha_A(y))} \tag{28}$$

$$= \sum_{y\in\mathsf{X}_A,\,\boldsymbol{d}_A(y)>0}\;\sum_{\ell=0}^{\boldsymbol{d}_A(y)-1} \ln(\alpha_A(y)+\ell), \tag{29}$$

for $A \subseteq N$, gives a (non-standardized) data vector relative to LML_α.

Proof. The first formula (28) follows from Proposition 8 by repeating the consideration in the proof of Corollary 9. The formula (29) then follows from the basic facts about Gamma functions from § A: if $\boldsymbol{d}_A(y) = 0$ then the ratio $\frac{\Gamma(\alpha_A(y)+\boldsymbol{d}_A(y))}{\Gamma(\alpha_A(y))}$ is 1 and $\ln 1 = 0$, if $\boldsymbol{d}_A(y) > 0$ then formula (34) is applied. □

Now, the uniquely determined standardized data vector $t_D^{\mathsf{LML},\alpha}$ can be obtained by standardization (see Lemma 8.7 in [27]):

$$t_D^{\mathsf{LML},\alpha}(A) = \check{t}_D^{\mathsf{LML},\alpha}(A) - \sum_{i\in A} \check{t}_D^{\mathsf{LML},\alpha}(\{i\}) + (|A|-1)\cdot \check{t}_D^{\mathsf{LML},\alpha}(\emptyset) \quad \text{for every } A \subseteq N.$$

More specifically, by (28) we have for any $A \subseteq N$, $|A| \geq 2$:

$$\begin{aligned} t_D^{\mathsf{LML},\alpha}(A) &= \sum_{y\in \mathsf{X}_A} \ln \frac{\Gamma(\alpha_A(y) + \boldsymbol{d}_A(y))}{\Gamma(\alpha_A(y))} \\ &\quad - \sum_{i\in A}\sum_{z\in \mathsf{X}_i} \ln \frac{\Gamma(\alpha_{\{i\}}(z) + \boldsymbol{d}_{\{i\}}(z))}{\Gamma(\alpha_{\{i\}}(z))} + (|A|-1)\cdot \ln \frac{\Gamma(\alpha_+ + d)}{\Gamma(\alpha_+)}, \end{aligned} \tag{30}$$

where $\alpha_+ \equiv \sum\{\alpha(x);\, x \in \mathsf{X}_N\}$. Of course, (28) can be replaced by (29) here.

Remark By an analogous consideration the saturating function can be expressed:

$$\begin{aligned} s_D^{\mathsf{LML},\alpha} &= \check{t}_D^{\mathsf{LML},\alpha}(N) - \check{t}_D^{\mathsf{LML},\alpha}(\emptyset) = \sum_{x\in \mathsf{X}_N} \ln \frac{\Gamma(\alpha(x) + \boldsymbol{d}(x))}{\Gamma(\alpha(x))} - \ln \frac{\Gamma(\alpha_+ + d)}{\Gamma(\alpha_+)} \\ &= \sum_{x\in \mathsf{X}_N,\, \boldsymbol{d}(x)>0} \sum_{\ell=0}^{\boldsymbol{d}(x)-1} \ln(\alpha(x)+\ell) - \sum_{\ell=0}^{d-1} \ln(\alpha_+ + \ell). \end{aligned}$$

A Gamma function

The following facts can be found in several handbooks of mathematics, e.g. in [4].

Definition 9 *Gamma function* Γ is defined on the interval $(0, +\infty)$ by the formula:

$$\forall\, \alpha > 0 \qquad \Gamma(\alpha) = \int_0^{+\infty} e^{-t}\cdot t^{\alpha-1}\, dt\ . \tag{31}$$

In particular,

- $\Gamma(\alpha) > 0$ for any $\alpha > 0$,
  because the function inside the integral in (31) is strictly positive on $(0, \infty)$,
- $\lim_{\alpha\to 0+} \Gamma(\alpha) = +\infty$
  because the limit is, in fact, the integral $\int_0^{+\infty} e^{-t}\cdot t^{-1}\, dt$, but $\lim_{t\to 0+} e^{-t} = 1$ and t^{-1} has infinite integral at 0: $\int_0^1 t^{-1}\, dt = +\infty$.

Two well-known values are

- $\Gamma(1) = 1$, and
- $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

The basic formula is as follows:

$$\forall\, \alpha > 0 \qquad \Gamma(\alpha+1) = \alpha\cdot\Gamma(\alpha)\,. \tag{32}$$

It implies well-known relation of Gamma function to the factorial:

$$\forall\, n \in \mathbb{N} \qquad \Gamma(n+1) = n!\ . \tag{33}$$

Indeed, this follows by the induction from (32) and $\Gamma(1) = 1$.

Another well-known formula relates Gamma funtion to Beta function:

$$\forall\, p, q > 0 \qquad B(p,q) = \frac{\Gamma(p)\cdot\Gamma(q)}{\Gamma(p+q)}\,,$$

where the Beta function B is defined by the formula $B(p,q) = \int_0^1 x^{p-1}\cdot(1-x)^{q-1}\,dx$.

The following formula is utilized in § 8 of this report:

Fact 1 $\forall\,\alpha > 0 \quad \forall\, d \in \mathbb{N}$

$$\ln\frac{\Gamma(\alpha+d)}{\Gamma(\alpha)} = \sum_{\ell=0}^{d-1} \ln(\alpha+\ell)\,. \tag{34}$$

Proof. By induction after d, for a fixed α. The induction hypothesis for $d = 1$ follows from (32): $\ln\frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} = \ln\frac{\alpha\cdot\Gamma(\alpha)}{\Gamma(\alpha)} = \ln\alpha = \sum_{\ell=0}^{1-1}\ln(\alpha+\ell)$. Note that $\Gamma(\alpha) > 0$ for $\alpha > 0$ makes this computation possible. The induction step for $d \geq 2$:

$$\begin{aligned}
\ln\frac{\Gamma(\alpha+d)}{\Gamma(\alpha)} &= \ln\frac{\Gamma(\alpha+d)}{\Gamma(\alpha+d-1)}\cdot\frac{\Gamma(\alpha+d-1)}{\Gamma(\alpha)} = \ln\frac{\Gamma(\alpha+d)}{\Gamma(\alpha+d-1)} + \ln\frac{\Gamma(\alpha+d-1)}{\Gamma(\alpha)} \\
&= \ln\frac{(\alpha+d-1)\cdot\Gamma(\alpha+d-1)}{\Gamma(\alpha+d-1)} + \ln\frac{\Gamma(\alpha+d-1)}{\Gamma(\alpha)} \\
&= \ln(\alpha+d-1) + \sum_{\ell=0}^{d-2}\ln(\alpha+\ell) = \sum_{\ell=0}^{d-1}\ln(\alpha+\ell)\,.
\end{aligned}$$

Indeed, since $d \geq 2$ and $\alpha > 0$ one has $\alpha + d - 1 > 0$ and $\Gamma(\alpha+d-1) > 0$, which allows one to make the step in the first line. Then (32) for $\alpha + d - 1$ is used in the second line and both cancelling of $\Gamma(\alpha+d-1)$ and the induction premise in the third line. □

B Uniformly distributed measures

B.1 Volume of a ball in the Euclidean space

It is a well-known fact that the volume (= the n-dimensional Lebesgue measure) of every ball in $\mathbb{R}^n$ with diameter $s > 0$ is $\kappa_n \cdot s^n$, where κ_n is a constant, only depending on the dimension n.[34] More specifically, the volume κ_n of every unit ball in $\mathbb{R}^n$ is given by:

$$\kappa_n = \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)} = \frac{2\cdot\pi^{\frac{n}{2}}}{n\cdot\Gamma(\frac{n}{2})}\,. \tag{35}$$

See, for example, Exercise 15.10 on pages 431-432 of [2].

Example 1 The well-known formulas say $\kappa_1 = 2$, $\kappa_2 = \pi$ and $\kappa_3 = \frac{4}{3}\cdot\pi$. However, one has $\kappa_4 = \frac{1}{2}\cdot\pi^2$. □

[34] Nevertheless, the value of the constant κ_n is not substantial for the later considerations in this report.

B.2 Isometries between affine spaces

Definition 10 (affine subspace of an Euclidean space)
Assume $r \geq 2$, denote $[r] \equiv \{1, \dots r\}$ and consider the Euclidean space $\mathbb{R}^{[r]}$. By an *affine subspace* of $\mathbb{R}^{[r]}$ of the dimension n, $1 \leq n \leq r$ is meant the set A of the form $A = x + L$, where $x \in \mathbb{R}^{[r]}$ and $L \subseteq \mathbb{R}^{[r]}$ is a linear subspace of the dimension n.[35]

An example of an affine subspace was mentioned in § 4.1:

$$A \,=\, \{\, (\theta_1, \dots, \theta_r)\,;\, \sum_{k=1}^{r} \theta_k = 1 \,\}\,.$$

Here, one can take $x = (1, 0, \dots, 0)$ and $L = \{\, (\theta_1, \dots, \theta_r)\,;\, \sum_{k=1}^{r} \theta_k \,=\, 0\,\}$. Thus, its dimension is $r - 1$. Every affine subspace can be viewed as a (separable) metric space endowed with the restriction of the Euclidean metric from $\mathbb{R}^{[r]}$:

$$\rho(x, y) \,=\, \sqrt{\sum_{k=1}^{r} (x_k - y_k)^2} \qquad \text{for } x, y \in A \subseteq \mathbb{R}^{[r]}\,.$$

Note that the metric on A is, in fact, determined by the set A itself through its identification as a subset of an Euclidean space; the dimension r of the Euclidean space in which A is embedded only has an auxiliary role. Below we show that every affine subspace of the dimension n is isometrical to $\mathbb{R}^{[n]}$ with n-dimensional Euclidean metric (Lemma 12).

Definition 11 (isometry)
Let (M, ρ) and (N, ς) be metric spaces. An *isometry* (between M and N) is a mapping $t : M \to N$ onto N which transfers the metric:

$$\forall\, x, y \in M \qquad \rho(x, y) \,=\, \varsigma(t(x), t(y))\,.$$

It is evident that an isometry is a one-to-one mapping (between M and N).[36] Moreover, the inverse mapping to an isometry is also an isometry. Every isometry naturally transfers metrical concepts like separability. One can also observe that isometry transfers Borel sets on M to Borel sets on N. Therefore, one can naturally transfer every Borel measure on M to (a Borel measure on) N.

We will utilize three basic ways to establish/construct an isometry between affine subspaces of the Euclidean space(s). These are:

A a *shift* (in $\mathbb{R}^{[r]}$),

B an *orthogonal transformation* (in $\mathbb{R}^{[r]}$),

C an *embedding* (of $\mathbb{R}^{[r]}$ into $\mathbb{R}^{[r']}$, $r' > r$).

[35] It is straightforward that the linear subspace L in $A = x + L$ is uniquely determined by A. Therefore, the dimension of A is uniquely defined, too.

[36] Indeed, $\forall\, x, y \in M \;\; t(x) = t(y) \;\Rightarrow\; \varsigma(t(x), t(y)) = 0 \;\Rightarrow\; \rho(x, y) = 0 \;\Rightarrow\; x = y$.

A Shift in the space $\mathbb{R}^{[r]}$

This is a transformation

$$y \in \mathbb{R}^{[r]} \;\longmapsto\; y + z \in \mathbb{R}^{[r]}\,,$$

where z is a fixed vector in $\mathbb{R}^{[r]}$. It transfers an affine space $A = x + L$ to the affine space $A' = (x + z) + L$. If we restrict this mapping to A then it is an isometry between A and A'.

Example 2 The set $A = \{\,(\theta_1, \ldots, \theta_r)\,;\ \sum_{k=1}^{r} \theta_k = 1\,\}$ is an affine subspace. One can choose the back shift with $z = (-1, 0, \ldots, 0)$ and trasform A by $y \mapsto y + z$ to the linear subspace $L = \{\,(\theta_1, \ldots, \theta_r)\,;\ \sum_{k=1}^{r} \theta_k = 0\,\}$. □

B Orthogonal transformations in $\mathbb{R}^{[r]}$

These are linear mappings/tranformations of $\mathbb{R}^{[r]}$ onto itself which preserve the angles between vectors. To construct them one can utilize the concept of an orthonormal basis (for $\mathbb{R}^{[r]}$), which has the meaning of a coordinate system.

Definition 12 (orthonormal basis)
By an *orthonormal basis* of the Euclidean space $\mathbb{R}^{[r]}$, $r \geq 1$ is understood a finite set of vectors $\mathcal{E} \subseteq \mathbb{R}^{[r]}$ such that two conditions hold:

- $\forall\, u, v \in \mathcal{E} \quad \langle u, v\rangle = \delta_{uv}$,
  where $\langle u, v\rangle \equiv \sum_{k=1}^{r} u_k \cdot v_k$ and δ_{uv} is the Dirac's delta symbol: $\delta_{uv} \equiv \begin{cases} 1 & \text{if } u = v, \\ 0 & \text{if } u \neq v, \end{cases}$
- $\forall\, x \in \mathbb{R}^{[r]} \quad \exists\, \alpha_u \in \mathbb{R}, u \in \mathcal{E}$ such that $x = \sum_{u\in\mathcal{E}} \alpha_u \cdot u\,.$

Note that the first condition can be formulated in this form: $\forall\, u, v \in \mathcal{E}, u \neq v$ one has $\langle u, v\rangle = 0$, which means the vectors are mutually orthogonal, and, $\forall\, u \in \mathcal{E}$ one has $\|u\|^2 \equiv \langle u, u\rangle = 1$, which means each vector $u \in \mathcal{E}$ has the length $\|u\| = 1$. The condition implies the elements of $\mathcal{E}$ are linearly independent.[37] Owing to the second condition, $\mathcal{E}$ forms a linear basis of $\mathbb{R}^{[r]}$, and, therefore, $|\mathcal{E}| = r$. Since it is a linear basis, for any $x \in \mathbb{R}^{[r]}$, the decomposition $x = \sum_{u\in\mathcal{E}} \alpha_u \cdot u$ has uniquely determined coefficients α_u. In particular, they can be regarded as the "coordinates" of x with respect to (a coordinate system) $\mathcal{E}$.

Lemma 11 *Given $r \geq 1$, $\mathcal{E} : u_1, \ldots u_r$ and $\mathcal{F} : v_1, \ldots v_r$ two ordered orthonormal bases of $\mathbb{R}^{[r]}$, there exists (just one) linear mapping $O : \mathbb{R}^{[r]} \to \mathbb{R}^{[r]}$ such that $\forall\, k = 1, \ldots, r$ $O(u_k) = v_k$. This mapping is an isometry of $\mathbb{R}^{[r]}$ onto itself.*

Proof. The first step is to observe that, for the chosen orderings $u_1, \ldots u_r$ and $v_1, \ldots v_r$, there exists an $r \times r$-matrix $\mathbb{O}$ such that $\forall\, k = 1, \ldots, r \quad \mathbb{O} \cdot u_k = v_k$.
For example, consider an auxiliary concept of the *standard orthonormal base* $e_1, \ldots, e_r$, that is, column vectors e_k given by $(e_k)_l = \delta_{kl}$ for $k, l = 1, \ldots, r$. There exists a matrix $\mathbb{W}$ with $\mathbb{W} \cdot e_k = v_k$ for $k = 1, \ldots, r$: it suffices to put v_k as the k-th column of $\mathbb{W}$. Analogously, there exists a matrix $\mathbb{V}$ with $\mathbb{V} \cdot e_k = u_k$ for $k = 1, \ldots, r$. Since its columns are linearly independent, the matrix $\mathbb{V}$ is regular (= invertible). In particular, $\mathbb{V}^{-1} \cdot u_k = e_k$ for $k = 1, \ldots, r$ and the matrix $\mathbb{O} \equiv \mathbb{W} \cdot \mathbb{V}^{-1}$ satisfies $\forall\, k = 1, \ldots, r \quad \mathbb{O} \cdot u_k = \mathbb{W} \cdot (\mathbb{V}^{-1} \cdot u_k) = \mathbb{W} \cdot e_k = v_k$.

[37] Indeed, $\sum_{u\in\mathcal{E}} \alpha_u \cdot u = 0$ implies $\forall\, v \in \mathcal{E} \quad 0 = \langle 0, v\rangle = \langle \sum_{u\in\mathcal{E}} \alpha_u \cdot u, v\rangle = \sum_{u\in\mathcal{E}} \alpha_u \cdot \langle u, v\rangle = \sum_{u\in\mathcal{E}} \alpha_u \cdot \delta_{uv} = \alpha_v$. Thus, $\forall\, v \in \mathcal{E}\ \alpha_v = 0$.

The desired linear mapping $O : \mathbb{R}^{[r]} \to \mathbb{R}^{[r]}$ can be introduced as the multiplication by this matrix: $O(x) \equiv \mathbb{O} \cdot x$ for $x \in \mathbb{R}^{[r]}$. The next observation is that the matrix $\mathbb{O}$ is *unitary*, which implies, one has $\|\mathbb{O} \cdot x\| = \|x\|$ for every $x \in \mathbb{R}^{[r]}$.

For example, this follows from Theorem 2.6 (on page 48) in the textbook [11]: one of equivalent definitions of the unitary matrix is the condition 5° from [11] requiring that there exists an orthonormal basis $u_1, \ldots, u_r$ such that $\mathbb{O} \cdot u_1, \ldots, \mathbb{O} \cdot u_r$ is an orthonormal basis, too. Thus, another equivalent definition of the unitary matrix the condition 2° implies $\forall\, x \in \mathbb{R}^{[r]} \;\; \|\mathbb{O} \cdot x\| = \|x\|$.

Therefore, $\forall\, x, y \in \mathbb{R}^{[r]}$ one has by linearity of the mapping $x \mapsto \mathbb{O} \cdot x$:

$$\rho(x, y) \;=\; \|x - y\| \;=\; \|\mathbb{O} \cdot (x - y)\| \;=\; \|\mathbb{O} \cdot x - \mathbb{O} \cdot y\| \;=\; \rho(\mathbb{O} \cdot x, \mathbb{O} \cdot y)\,,$$

which means O is an isometry of $\mathbb{R}^{[r]}$ onto itself.[38] The uniqueness of this linear mapping follows from the uniqueness of the matrix $\mathbb{O}$.

Indeed, if $\mathbb{O}_1$ and $\mathbb{O}_2$ are two matrices of that form then $\mathbb{O}_1 - \mathbb{O}_2$ satisfies $(\mathbb{O}_1 - \mathbb{O}_2) \cdot u_k = 0$ for every $k = 1, \ldots, r$, and this, since $u_1, \ldots, u_r$ is a linear basis of $\mathbb{R}^{[r]}$, implies $\mathbb{O}_1 - \mathbb{O}_2$ is the zero matrix. □

An isometric linear mapping O of $\mathbb{R}^{[r]}$ onto itself transfers (every) linear subspace L of $\mathbb{R}^{[r]}$ to a linear subspace of the same dimension. Therefore, it is an isometry of the corresponding metric spaces (≡ of L and its image $O(L)$).

Example 3 The linear subspace $L = \{\, (\theta_1, \ldots, \theta_r)\,;\; \sum_{k=1}^{r} \theta_k = 0 \,\}$ of $\mathbb{R}^{[r]}$ can be transferred by such an isometry to another linear subspace

$$K \;=\; \{\, (\theta_1, \ldots, \theta_r)\,;\; \theta_r = 0 \,\}\,.$$

To this end, it suffices to choose a suitable orthonormal basis $\mathcal{E} : u_1, \ldots, u_r$ such that the linear hull of $u_1, \ldots, u_{r-1}$ is just L. For example, one can put

$$u_k = (\underbrace{\frac{1}{\sqrt{k} \cdot \sqrt{k+1}}, \ldots, \frac{1}{\sqrt{k} \cdot \sqrt{k+1}}}_{k\text{–times}}, \frac{-\sqrt{k}}{\sqrt{k+1}}, 0, \ldots, 0) \qquad \text{for } k = 1, \ldots, r-1\,,$$

and $u_r = (\frac{1}{\sqrt{r}}, \ldots, \frac{1}{\sqrt{r}})$. In the place of $\mathcal{F}$ one can choose the standard orthonormal basis $e_1, \ldots, e_r$, which has the property that the linear hull of $e_1, \ldots, e_{r-1}$ is just K. Then Lemma 11 can be applied to get an isometric linear mapping $O : \mathbb{R}^{[r]} \to \mathbb{R}^{[r]}$: it tranfers L to K. □

C Embedding of $\mathbb{R}^{[r]}$ into $\mathbb{R}^{[r']}$, $r' > r$.

This is a mapping which identifies every vector of the lower dimension r with a vector of the higher dimension r' by "adding" zero components. More specifically:

$$(x_1, \ldots, x_r) \;\longmapsto\; (y_1, \ldots, y_{r'}) \qquad \text{where } \; y_k = \begin{cases} x_k & \text{for } k = 1, \ldots, r, \\ 0 & \text{for } k = r+1, \ldots, r'. \end{cases}$$

It is straightforward from the definition of the metrics that it isometrically transfers $\mathbb{R}^{[r]}$ to the linear subspace $\{(y_1, \ldots, y_{r'});\; y_{r+1} = \ldots = y_{r'} = 0\}$ of $\mathbb{R}^{[r']}$.[39]

[38] Another equivalent definition of an unitary matrix is – see the condition 3° in Theorem 2.6 of [11] – is that $\forall\, x, y \in \mathbb{R}^{[r]}$ one has $\langle \mathbb{O} \cdot x, \mathbb{O} \cdot y \rangle = \langle x, y \rangle$, which means the mapping $x \mapsto \mathbb{O} \cdot x$ preserves the scalar products of vectors (= the angles between vectors).

[39] Indeed, the distance of two vectors u, v in both spaces is $\sqrt{\sum_{k=1}^{r} (u_k - v_k)^2}$.

Example 4 The image of $\mathbb{R}^{[r-1]}$, $r \geq 2$ by the embedding

$$\iota : (\theta_1, \ldots, \theta_{r-1}) \longmapsto (\theta_1, \ldots, \theta_{r-1}, 0)$$

is the (above-mentioned) linear subspace $K = \{ (\theta_1, \ldots, \theta_r) ;\ \theta_r = 0 \}$ of $\mathbb{R}^{[r]}$. The inverse mapping to ι, therefore, transfers isometrically K onto $\mathbb{R}^{[r-1]}$. □

Lemma 12 *Let A be an affine subspace of $\mathbb{R}^{[r]}$, $r \geq 2$ of the dimension $n \geq 1$. Then A is isometrically isomorphic to the Euclidean space $\mathbb{R}^{[n]}$.*

Proof. One can utilize the above mentioned constructions A-C of isometric mappings between affine subspaces. The first step is that $A = x + L$ is isometrically transferred by a shift to the (uniquely determined) linear subspace $L \subseteq \mathbb{R}^{[r]}$ of the dimension n.

The second step is an orthogonal transformation in $\mathbb{R}^{[r]}$ which isometrically transfers L to the linear subspace

$$K \ = \ \{ (\theta_1, \ldots, \theta_r) ;\ \theta_{n+1} = \ldots = \theta_r = 0 \} .$$

This mapping can be constructed by Lemma 11.
Indeed, one first chooses a linear basis $b_1, \ldots, b_n$ of L and completes it to a linear basis $b_1, \ldots, b_r$ of the whole space $\mathbb{R}^{[r]}$. Then one can apply well-known Gram-Schmidt process (for orthogonalizing) to it (see e.g. page 51 in [11]). The result is an (ordered) orthonormal base $\mathcal{E} : u_1, \ldots, u_r$ such that the linear hull of $u_1, \ldots, u_n$ is just L. Then one can choose for $\mathcal{F}$ the standard orthonormal base $e_1, \ldots, e_r$ of $\mathbb{R}^{[r]}$. It has the property that the linear hull of $e_1, \ldots, e_n$ is just K. Then Lemma 11 is applied to get a linear mapping O with $O(u_k) = e_k$ for $k = 1, \ldots, r$, which is an isometry.

The third step is to observe that K is isometrically isomorphic to $\mathbb{R}^{[n]}$: for this purpose, one can use the inverse mapping to the embedding of $\mathbb{R}^{[n]}$ into $\mathbb{R}^{[r]}$. Of course, the composition of these three isometries is again an isometry. □

B.3 Proper Lebesgue measure on an affine subspace

The above-mentioned Lemma 12 is the first step to introduce the concept of a proper Lebesgue measure on an affine subspace of an Euclidean space. The second step is the next definition.

Definition 13 (uniformly distributed measure)
Let (M, ρ) be a separable metric space. Given $x \in M$ and $s > 0$, let us denote by

$$U(x, s) \ = \ \{ y \in M ;\ \rho(x, y) \leq s \}$$

the *closed ball* around x with diameter s. A Borel measure μ on (M, ρ) will be called *locally finite* if $\forall\, x \in M \ \ \exists\, s > 0$ with $\mu(U(x, s)) < \infty$. A locally finite Borel measure μ on (M, ρ) will be called *uniformly distributed* if

$$\forall\, x, y \in M \quad \forall\, s > 0 \qquad \mu(U(x, s)) = \mu(U(y, s)) .$$

It is straightforward that an isometry between metric spaces M and N tranfers a uniformly distributed measure (on M) to a uniformly distributed measure (on N).[40] The basic observation concerning uniformly distributed measures is this:

[40] This is a trival consequence of the fact that an isometry transfers a ball with diameter $s > 0$ to a ball with diameter s.

Lemma 13 *A non-zero uniformly distributed measure on a separable metric space M is uniquely determined up to a positive multiple. In particular, if $s > 0$ is such a diameter that measure of the ball with diameter s is non-zero and finite, then uniformly distributed measure on M is determined uniquely by the measure of this ball.*

Proof. Note that this uniqueness result was already given in [8], in the context of a locally compact Hausdorff topological space with uniform structure. Lemma 13 follows from the results of [26]:
Specifically, Consequence V2 on p. 61 in [26] says this: if there are two uniformly distributed measures μ and ν on a separable metric space and $\mu \neq 0$ then there exists $t \geq 0$ such that $\nu = t \cdot \mu$. Thus, if $\nu \neq 0$ then $t > 0$. This observation implies that whenever $s > 0$ exists such that $\mu(U(x,s)) < \infty$ for some $x \in M$ (which means, since μ is uniformly distributed, for every $x \in M$) then $\nu(U(x,s)) < \infty$ for any other non-zero uniformly distributed measure ν on M. Therefore, if $0 < \mu(U(x,s)) < \infty$ and $\mu(U(x,s)) = \nu(U(x,s))$ for some $x \in M$ then the relation $\nu = t \cdot \mu$ gives $t = 1$. Hence, $\nu = \mu$. □

Now, one can introduce the concept of a proper Lebesgue measure on an affine subspace.

Definition 14 (proper Lebesgue measure on an affine subspace)
By a *proper Lebesgue measure* on an affine subspace $A \subseteq \mathbb{R}^{[r]}$, $r \geq 2$ of the dimension $n \geq 1$ will be meant (necessarily non-zero) uniformly distributed Borel measure λ_A on A such that the measure of the unit ball in A is the same as in the Euclidean space $\mathbb{R}^{[n]}$, that means,

$$\lambda_A(U_A(x,1)) \,=\, \kappa_n \qquad \text{for every } x \in A\,,$$

where $U_A(x,s)$ denotes the closed ball in A with diameter s.

Finally, the previous facts allow one to get the basic existence and uniqueness result:

Proposition 14 *On every affine subspace A (of the Euclidean space) of the dimension $n \geq 1$ there exists a proper Lebesgue measure λ_A, and is determined uniquely. It satisfies the following formula*

$$\lambda_A(U_A(x,s)) \,=\, \kappa_n \cdot s^n \qquad \text{for every } x \in A,\ s > 0\,. \tag{36}$$

Proof. By Lemma 12, A is isometrically isomorphic to $\mathbb{R}^{[n]}$ and the standard Lebesgue measure on $\mathbb{R}^{[n]}$ satisfies the requirement that the measure of the ball with diameter $s > 0$ is $\kappa_n \cdot s^n$. This measure is transferred by the isometry to A and one gets in this way a (locally finite) uniformly distributed measure on A such that the measure of the unit ball is κ_n. Therefore, the proper Lebesgue measure on A exists and its uniqueness follows from Lemma 13: for $s = 1$ the measure of the ball with diameter s is non-zero and finite. The formula (36) for arbitrary $s > 0$ is transferred from $\mathbb{R}^{[n]}$ by the isometry. □

It follows from the definition that an isometry between affine spaces transfers proper Lebesgue measures on themselves.

B.3.1 Lebesgue measure of cuboids

The proper Lebesgue measure on a linear subspace L of an Euclidean space can equivalently be introduced as the (uniformly distributed) measure on L which ascribes the value one to every unit cube in L. Of course, the concept of a cube, respectively of a cuboid, in a linear subspace is relative to a coordinate system.

Definition 15 (cuboid in a linear subspace)
Given $r \geq 2$ and a linear subspace L of $\mathbb{R}^{[r]}$ of the dimension $n \geq 1$, let $\mathcal{E}$ be an orthonormal basis of L. Then by the *$\mathcal{E}$-cuboid* determined by the parameters $a_u < b_u$, $u \in \mathcal{E}$ will be meant the set of those vectors in L whose coordinates with respect to $\mathcal{E}$ fall within those limits:

$$\prod_{u\in\mathcal{E}}(a_u, b_u] \equiv \{x \in L; \text{ if } x = \sum_{u\in\mathcal{E}} \alpha_u \cdot u \text{ then } a_u < \alpha_u \leq b_u \text{ for } u \in \mathcal{E}\}.$$

Example 5 If $L = \mathbb{R}^{[r]}$ and $\mathcal{E}$ is the standard orthonormal basis $e_1, \ldots, e_r$ of L then $\mathcal{E}$-cuboid is nothing but the usual classic cuboid, that is, the Cartesian product of correspoding intervals: $\prod_{u\in\mathcal{E}}(a_u, b_u] \equiv \prod_{i=1}^{r}(a_i, b_i]$. □

To prove the desired formula for (proper) Lebesgue measure of a cuboid we need an auxiliary concept, well-known from the textbooks on probability theory, e.g. [25]:

Definition 16 (σ-additive system)
Let $\mathcal{S}$ be a system of subsets of a non-empty set X. It will be named *σ-additive* if

- it is closed under disjoint finite union: $\forall A, B \in \mathcal{S} \quad A \cap B = \emptyset \quad \Rightarrow \quad A \cup B \in \mathcal{S}$,
- under proper difference: $\forall A, B \in \mathcal{S} \quad A \subseteq B \quad \Rightarrow \quad B \setminus A \in \mathcal{S}$, and
- under monotone countable union: $\forall \{A_n\} \subseteq \mathcal{S} \quad A_n \subseteq A_{n+1}, n \in \mathbb{N} \quad \Rightarrow \quad \bigcup_{n=1}^{\infty} A_n \in \mathcal{S}$.

It is evident that the collection of those (measurable) sets on which two finite non-negative measures equal each other is a σ-additive system. Well-known fact (see e.g. Statement I.15 on page 35 of [25]) is that whenever $\mathcal{L}$ is a system of subsets of X closed under finite intersection ($\equiv \forall A, B \in \mathcal{L} \quad A \cap B \in \mathcal{L}$) and containing the basic set ($\equiv X \in \mathcal{L}$) then the least σ-additive system containing $\mathcal{L}$ is the σ-algebra generated by $\mathcal{L}$.

This implies that, for two (σ-finite) Borel measures μ and ν on a metric space M, to show $\mu = \nu$ it suffices to show that μ and ν are finite and equal on a system $\mathcal{K}$ of sets,

- which is closed under (finite) intersection,
- M is the monotone countable union of some elements in $\mathcal{K}$, and
- the least σ-additive system containing $\mathcal{K}$ is the Borel σ-algebra on M.

Indeed, since $M = \bigcup_{n\in\mathbb{N}} S_n$ where $S_n \in \mathcal{K}$, $S_n \subseteq S_{n+1}$ for $n \in \mathbb{N}$, it suffices to verify for every $n \in \mathbb{N}$ that $\mu = \nu$ on Borel subsets of S_n. This is because $\mu(A) = \lim_{n\to\infty} \mu(A \cap S_n) = \lim_{n\to\infty} \nu(A \cap S_n) = \nu(A)$ for every Borel set $A \subseteq M$. One can consider the class $\mathcal{L} \equiv \{K \cap S_n; K \in \mathcal{K}\} \subseteq \mathcal{K}$. Then $\mu = \nu$ on $\mathcal{L}$ and $\nu(S_n) = \mu(S_n) < \infty$ by the assumption. Therefore, $\mu = \nu$ on the least σ-additive system containing $\mathcal{L}$, that is, on the σ-algebra $\sigma(\mathcal{L})$ by the above-mentioned fact applied to $X = S_n$. However, $\sigma(\mathcal{L})$ is the class of Borel sets in S_n.

Lemma 15 *Given $r \geq 2$, let L be a linear subspace of $\mathbb{R}^{[r]}$ of the dimension $n \geq 1$ and $\mathcal{E}$ an orthonormal basis of L. Then*

(i) *The class $\mathcal{K}_\mathcal{E}$ of all $\mathcal{E}$-cuboids in L is closed under (finite) intersection. The least σ-additive system of subsets of L containing $\mathcal{K}_\mathcal{E}$ is the Borel σ-algebra on L.*

(ii) *The proper Lebesgue measure (on L) of $\mathcal{E}$-cuboids is determined by the formula*

$$\lambda_L\,(\,\prod_{u\in\mathcal{E}}(a_u,b_u]\,)\;=\;\prod_{u\in\mathcal{E}}(b_u-a_u)\qquad \textit{whenever } a_u<b_u, u\in\mathcal{E}\,. \tag{37}$$

Proof. Let us order the elements of $\mathcal{E}$ into a sequence $u_1,\dots,u_n$ and then complete it to an orthonormal basis $u_1,\dots,u_r$ of the whole $\mathbb{R}^{[r]}$. Then consider the standard orthonormal basis $e_1,\dots,e_r$ in $\mathbb{R}^{[r]}$. By Lemma 11 construct a linear isometric mapping O of $\mathbb{R}^{[r]}$ onto itself such that $O(u_k)=e_k$ for $k=1,\dots,r$. This one-to-one mapping transfers L onto $K\equiv\{(\theta_1,\dots,\theta_r);\ \theta_{n+1}=\dots=\theta_r=0\,\}$. The linear subspace K can be identified by the inverse ι^{-1} (of the correspoding embedding) with $\mathbb{R}^{[n]}$. It is clear that $U\equiv\iota^{-1}\circ O$ is a one-to-one mapping from L to $\mathbb{R}^{[n]}$, saves set operations and transfers $\mathcal{E}$-cuboids to classic cuboids in $\mathbb{R}^{[n]}$.

Indeed, if $u=\sum_{k=1}^r\alpha_k\cdot u_k\in L$ for some $\alpha_k\in\mathbb{R}$ then $\alpha_k=0$ for $k=n+1,\dots,r$ and $O(u)=O(\sum_{k=1}^n\alpha_k\cdot u_k)=\sum_{k=1}^n\alpha_k\cdot O(u_k)=\sum_{k=1}^n\alpha_k\cdot e_k$ and $U(u)=\sum_{k=1}^n\alpha_k\cdot\hat{e}_k$, where $\hat{e}_1,\dots,\hat{e}_n$ is the standard orthonormal basis for $\mathbb{R}^{[n]}$.

Since the class of classic cuboids is closed under (finite) intersection, the same is true for the class of $\mathcal{E}$-cuboids. Because U saves set operations, it transfers the least σ-additive system containing $\mathcal{K}_\mathcal{E}$ to the least σ-additive system containing classic cuboids. This is, however, the Borel σ-algebra on $\mathbb{R}^{[n]}$, which, by measurability of U^{-1}, is transferred to the Borel σ-algebra on L. This implies the condition (i).

Since U is an isometry of L and $\mathbb{R}^{[n]}$, it transfers the (proper) Lebesgue measure on L to the standard Lebesgue measure on $\mathbb{R}^{[n]}$. Therefore, λ_L measure of the $\mathcal{E}$-cuboid $\prod_{u\in\mathcal{E}}(a_u,b_u]$ is the n-dimensional Lebesgue measure of its image, that is, of $\prod_{i=1}^n(a_{u_i},b_{u_i}]$. This equals to $\prod_{i=1}^n(b_{u_i}-a_{u_i})=\prod_{u\in\mathcal{E}}(b_u-a_u)$. This gives the condition (ii). □

The consequence of the formula (37) is that the proper Lebesgue measure of a unit cube, that is, of a cuboid $\prod_{u\in\mathcal{E}}(a_u,b_u]$ with $b_u-a_u=1$ for $u\in\mathcal{E}$, is 1. Moreover, it follows from the arguments above Lemma 15 that the unique Borel measure satisfying the formula (37) is the proper Lebesgue measure on L.

B.4 Lifting less-dimensional Lebesgue measure

In this section, we introduce a special lifting transformation from the Euclidean space to a certain affine subspace in a higher-dimensional space. Then we show it transforms the Lebesgue measure to a multiple of the proper Lebesgue measure on the affine subspace.

Definition 17 (lifting transformation)
Assume $r\geq 2$, denote $[r]\equiv\{1,\dots r\}$ and consider $l\in[r]$. By a *lifting mapping* from $\mathbb{R}^{[r]\setminus\{l\}}$ to the affine space

$$A\;=\;\{\,(\theta_1,\dots,\theta_r)\,;\,\sum_{k=1}^r\theta_k=1\,\}$$

will be meant the mapping $\mathcal{L}_l:\mathbb{R}^{[r]\setminus\{l\}}\to A$ defined as follows:

$$\mathcal{L}_l:\ [\eta_k]_{k\in[r]\setminus\{l\}}\quad\longmapsto\quad[\theta_k]_{k\in[r]}\qquad\text{where}\quad\begin{array}{l}\theta_k=\eta_k\quad\text{for } k\in[r]\setminus\{l\}\,,\\ \theta_l=1-\sum_{k\in[r]\setminus\{l\}}\eta_k\,.\end{array}$$

The lifting mapping $\mathcal{L}_l$ from $\mathbb{R}^{[r]\setminus\{l\}}$ to A is not (a multiple of) an isometry, but one can imagine it as the composition of three mappings, two of which are isometries.

- The first one is the embedding of $\mathbb{R}^{[r]\setminus\{l\}}$ into $\mathbb{R}^{[r]}$, whose image is the linear subspace $K_l = \{(\eta_1,\dots,\eta_r);\ \eta_l = 0\} \subseteq \mathbb{R}^{[r]}$ given by

$$\iota_l :\ [\eta_k]_{k\in[r]\setminus\{l\}} \quad\longmapsto\quad [\hat\eta_k]_{k\in[r]} \qquad \text{where} \quad \begin{array}{l}\hat\eta_k = \eta_k \quad \text{for } k\in[r]\setminus\{l\}\,,\\ \hat\eta_l = 0\,.\end{array}$$

- The second in a linear transformation $\mathcal{T}_l : K_l \to L$, where $L = \{\,(\theta_1,\dots,\theta_r)\,;\ \sum_{k=1}^r \theta_k = 0\,\}$, defined by the relation

$$\mathcal{T}_l :\ [\hat\eta_k]_{k\in[r]} \quad\longmapsto\quad [\hat\theta_k]_{k\in[r]} \qquad \text{where} \quad \begin{array}{l}\hat\theta_k = \hat\eta_k \quad \text{for } k\in[r]\setminus\{l\}\,,\\ \hat\theta_l = -\sum_{k\in[r]\setminus\{l\}} \hat\eta_k\,.\end{array}$$

  It transforms K_l onto L.

- The third mapping is a shift in $\mathbb{R}^{[r]}$ given by

$$\mathcal{S}_l :\ [\hat\theta_k]_{k\in[r]} \quad\longmapsto\quad [\hat\theta_k]_{k\in[r]} + [z_k^l]_{k\in[r]} \equiv [\theta_k]_{k\in[r]} \qquad \text{where}\ \ z_k^l = \delta_{kl} \text{ for } k\in[r]\,.$$

  It transforms L onto A.

The aim is to show that lifting transforms the Lebesgue measure to its certain multiple. The basic step to show this is to verify an analogous fact for the linear transformation $\mathcal{T}_l$.

Lemma 16 *Let $r \geq 2$, $[r] \equiv \{1,\dots r\}$ and $l \in [r]$. Then the image of proper Lebesgue measure on K_l by the linear transformation $\mathcal{T}_l$ is the $\frac{1}{\sqrt{r}}$-multiple of the proper Lebesgue measure on L:*

$$\lambda_{K_l} \circ (\mathcal{T}_l)^{-1} = \frac{1}{\sqrt{r}} \cdot \lambda_L\,.$$

Proof. Let us consider a linear subspace $K \equiv K_l \cap L$ of $\mathbb{R}^{[r]}$, that is,

$$K \;=\; \{\,[\theta_k]_{k\in[r]}\,;\ \theta_l = 0 \ \&\ \sum_{k\in[r]} \theta_k = 0\,\}.$$

Then choose an orthonormal basis $\mathcal{E}$ of K and introduce two vectors $u, v \in \mathbb{R}^{[r]}$:

$$v_k = \frac{1}{\sqrt{r-1}} \ \text{ for } k\in[r]\setminus\{l\}, \quad v_l = 0$$
$$u_k = \frac{1}{\sqrt{r-1}\cdot\sqrt{r}} \ \text{ for } k\in[r]\setminus\{l\}, \quad u_l = -\frac{\sqrt{r-1}}{\sqrt{r}}\,.$$

Clearly, $v \in K_l$, $u \in L$, $\|v\| = \|u\| = 1$. It is easy to see that $\mathcal{E}\cup\{v\}$ is an orthonormal basis of K_l and $\mathcal{E}\cup\{u\}$ an orthonormal basis of L.

Indeed, both v and u are perpendicular to all elements of K.

Let us observe that the transformation $\mathcal{T}_l$ can equivalently be introduced as follows:

$$\text{If } x\in K_l \quad \text{with} \quad x = \sum_{w\in\mathcal{E}\cup\{v\}} \alpha_w \cdot w \quad \text{and} \quad \mathcal{T}_l(x) = \sum_{w\in\mathcal{E}\cup\{u\}} \beta_w \cdot w$$
$$\text{then} \quad \forall\, w\in\mathcal{E} \quad \beta_w = \alpha_w \quad \& \quad \beta_u = \sqrt{r}\cdot\alpha_v\,.$$

Indeed, since $\mathcal{E}\cup\{v\}$ is an orthonormal basis of K_l the corresponding coefficients can be obtained as the scalar products:

$$\forall\,\bar{w}\in\mathcal{E}\cup\{v\}\quad \langle x,\bar{w}\rangle=\langle\sum_w \alpha_w\cdot w,\bar{w}\rangle=\sum_w \alpha_w\cdot\langle w,\bar{w}\rangle=\sum_w \alpha_w\cdot\delta_{w\bar{w}}=\alpha_{\bar{w}}\,.$$

Analogously, $\forall\,\bar{w}\in\mathcal{E}\cup\{u\}$ one has $\langle\mathcal{T}_l(x),\bar{w}\rangle=\beta_{\bar{w}}$ because $\mathcal{E}\cup\{u\}$ is an orthonormal basis of L and $\mathcal{T}_l(x)\in L$. Thus, for every $w\in\mathcal{E}$ we write

$$\alpha_w-\beta_w=\langle x,w\rangle-\langle\mathcal{T}_l(x),w\rangle=\langle x-\mathcal{T}_l(x),w\rangle=0\,,$$

where the last equality follows from the definition of $\mathcal{T}_l$: one has $(\mathcal{T}_l(x))_k=x_k$ for $k\in[r]\setminus\{l\}$ and $w_l=0$ for every $w\in\mathcal{E}\subseteq K$. Then we write analogously, owing to the definition of v:

$$\alpha_v=\langle x,v\rangle=\frac{1}{\sqrt{r-1}}\cdot\sum_{k\in[r]\setminus\{l\}}x_k\,,$$

and, owing to the definition $\mathcal{T}_l(x)$ and u:

$$\begin{aligned}\beta_u &= \langle\mathcal{T}_l(x),u\rangle=\frac{1}{\sqrt{r-1}\cdot\sqrt{r}}\cdot\sum_{k\in[r]\setminus\{l\}}x_k\;+\;(\frac{-\sqrt{r-1}}{\sqrt{r}})\cdot(-\sum_{k\in[r]\setminus\{l\}}x_k)\\ &= (\frac{1}{\sqrt{r-1}\cdot\sqrt{r}}+\frac{\sqrt{r-1}}{\sqrt{r}})\cdot\sum_{k\in[r]\setminus\{l\}}x_k=\frac{\sqrt{r}}{\sqrt{r-1}}\cdot\sum_{k\in[r]\setminus\{l\}}x_k\,,\end{aligned}$$

and, therefore, $\beta_u=\sqrt{r}\cdot\alpha_v$.
Now, the observation above implies that each $\mathcal{E}\cup\{v\}$-cuboid of the form $\prod_{w\in\mathcal{E}\cup\{v\}}(a_w,b_w]$ is by the transformation $\mathcal{T}_l$ transferred to the $\mathcal{E}\cup\{u\}$-cuboid of the form $\prod_{w\in\mathcal{E}\cup\{u\}}(a_w,b_w]$, where $a_u=\sqrt{r}\cdot a_v$ and $b_u=\sqrt{r}\cdot b_v$. Thus, by Lemma 15(ii) the proper Lebesgue measure of the former cuboid is

$$\lambda_{K_l}(\prod_{w\in\mathcal{E}\cup\{v\}}(a_w,b_w])=(b_v-a_v)\cdot\prod_{w\in\mathcal{E}}(b_w-a_w)\,,$$

and the proper Lebesgue measure of the latter one is

$$\lambda_L(\prod_{w\in\mathcal{E}\cup\{u\}}(a_w,b_w])=(b_u-a_u)\cdot\prod_{w\in\mathcal{E}}(b_w-a_w)=\sqrt{r}\cdot(b_v-a_v)\cdot\prod_{w\in\mathcal{E}}(b_w-a_w)\,.$$

In particular, $\forall\,A\in\mathcal{K}_{\mathcal{E}\cup\{v\}}$ one has $\lambda_L(\mathcal{T}_l(A))=\sqrt{r}\cdot\lambda_{K_l}(A)$, which means $\forall\,B\in\mathcal{K}_{\mathcal{E}\cup\{u\}}$ one has $\lambda_L(B)=\sqrt{r}\cdot\lambda_{K_l}((\mathcal{T}_l)^{-1}(B))$, that is, $\frac{1}{\sqrt{r}}\cdot\lambda_L(B)=\lambda_{K_l}\circ(\mathcal{T}_l)^{-1}(B)$. Thus, the equality $\frac{1}{\sqrt{r}}\cdot\lambda_L=\lambda_{K_l}\circ(\mathcal{T}_l)^{-1}$ holds for all $\mathcal{E}\cup\{u\}$-cuboids and both measures are finite on those cuboids. By Lemma 15(i) the smallest σ-additive system containing $\mathcal{K}_{\mathcal{E}\cup\{u\}}$ is the Borel σ-algebra on L. That's why the equality extends to all Borel set – see the arguments above Lemma 15. □

Now, the desired statement can easily be obtained.

Proposition 17 *Assume $r\geq 2$, put $[r]\equiv\{1,\dots r\}$ and consider $l\in[r]$. Then the image of the (standard) Lebesgue measure on the space $\mathbb{R}^{[r]\setminus\{l\}}$ by the lifting mapping $\mathcal{L}_l$ is the $\frac{1}{\sqrt{r}}$-multiple of the proper Lebesgue measure λ_A on A. In particular, the image of the Lebesgue measure on $\mathbb{R}^{[r]\setminus\{l\}}$ by $\mathcal{L}_l$ does not depend on the choice of l!*

Proof. This follows directly from Lemma 16 and the fact that $\mathcal{L}_l$ is the compositions of two isometries and the linear transformation $\mathcal{T}_l$. Indeed, isometries between affine spaces transfer proper Lebesgue measures onto themselves. □

References

[1] J. Anděl. *Mathematical Statistics*. (in Czech) SNTL (Prague) 1985.

[2] T.M. Apostol. *Mathematical Analysis*. Addison-Wesley (Reading) 1974.

[3] S.A. Andersson, D. Madigan and M.D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics* 25 (1997) 505-541.

[4] H.-J. Bartsch. *Alle Rechte Vorbehalten*. Carl Hanser Verlag 1994. Czech translation: *Matematické vzorce*. Mladá fronta 1996.

[5] R.R. Bouckaert. Bayesian belief networks: from construction to evidence. PhD thesis, University of Utrecht 1995.

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

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

[8] J.P.R. Christensen. On some measures analogous to Haar measure. *Mathematica Scandinavica* 26 (1970) 103-106.

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

[10] A.P. Dawid and S.L. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics* 21 (1993) 1272-1317.

[11] M. Fiedler. *Special Matrices and their Use in Numerical Mathematics*. (in Czech) SNTL (Prague) 1981.

[12] J.-P. Florens, M. Mouchart and J.-M. Rolin. *Elements of Bayesian Statistics*. Marcel Dekker 1990.

[13] D. Geiger, D. Heckerman, H. King and C. Meek. Stratified exponetial families: graphical models and model selection. *The Annals of Statistics* 29 (2001) 505-529.

[14] I.J. Good. *The Estimation of Probabilities*. MIT Press 1965.

[15] D.M.A. Haughton. On the choice of a model to fit data from an exponential family. *The Annals of Statistics* 16 (1988) 342-355.

[16] D. Heckerman. A tutorial on learning Bayesian networks. Technical report MSR-TR-95-06, Microsoft Research, Redmond, March 1995.

[17] D. Heckerman, D. Geiger and D.M. Chickering. Learing Bayesian networks: the combination of knowledge and statistical data. *Machine Learning* 20 (1995) 194-243.

[18] T. Kočka. Graphical models: learning and application. PhD thesis, University of Economics Prague 2001.

[19] S.L. Lauritzen. *Graphical Models*. Clarendon Press 1996.

[20] E. Lehman. *Testing Statistical Hypotheses*. Russian translation: Nauka 1979.

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

[22] R.E. Neapolitan. *Learning Bayesian Networks*. Pearson Prentice Hall 2004.

[23] G. Schwarz. Estimation the dimension of a model. *The Annals of Statistics* 6 (1978) 461-464.

[24] D.J. Spiegelhalter and S.L. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks* 20 (1990) 579-605.

[25] J. Štěpán. *Probability Theory: Mathematical Foundations.* (in Czech) Academia (Prague) 1987.

[26] M. Studený. The differentiation of measures in metric spaces. (in Czech), graduate diploma thesis, Faculty of Mathematics and Physics, Charles University, Prague 1981.

[27] M. Studený. *Probabilistic Conditional Independence Structures*. Springer-Verlag 2005.

[28] M. Studený and J. Vomlel. A geometric approach to learning BN structures. In *Proceedings the 4th European Workshop on Probabilistic Graphical Models* (M. Jaeger and T.D. Nielsen eds.), Hirtshals, Denmark, September 17-19, 2008, pp. 281-288.