Ludwig-Maximilians University

Seminar Paper

Introduction to continuous and hybrid Bayesian networks.

Seminar: (Imprecise) Probabilistic Graphical Models

Author:

Joanna Ficek

Supervisor:

M. Sc. Paul Fink

March 4, 2016

Contents

This seminar paper was written in LaTeX.

Abstract

Bayesian network is a probability framework able to handle many variables and incorporate uncertainty caused by partial knowledge or noisy observations. This probabilistic graphical model has been successfully applied in various fields including genetics and information theory. The most commonly used models are discrete Bayesian networks, however many real-life situations involve continuous variables (such as temperature or gene expression measurements). Both, continuous and hybrid networks require some additional machinery and care while applying algorithms known from the discrete case. The solutions usually applied are discretization of all the continuous quantities or approximation with the Gaussian distribution. Moreover, the use of the exponential form of a distribution may also substantially facilitate some operations. In case of unknown or very complex true distributions the notions of entropy and relative entropy play an important role and provide tools for finding the best approximation. One good solution in such situations is projecting the empirical distribution onto a desired family of distributions.

This paper will serve as an introduction to continuous and hybrid Bayesian networks and the mathematical theory underlying the listed methods and solutions will be presented. Many issues will only be tackled and hence, the interested reader will be suggested an appropriate source of information.

1 Introduction

Bayesian networks have recently received a lot of attention due to combination of several properties. They successfully deal with noisy or missing observations and they allow for incorporation of prior knowledge (e.g. expert's opinion, information from previous studies). Moreover, graphical representation naturally mirrors relationships between variables and facilitates adding data with the time. The applications of Bayesian networks can be found in various areas of knowledge. Among others:

- in the information theory (e.g. pattern classification) [Frey B.J., 1998, p.58];
- in medicine (supporting decision-making under uncertainty) [Nikovski D., 2000];
- in genetics (e.g. constructing gene regulatory networks; genetic analyses related to pedigree) [Friedman et al., 2000][Lauritzen S. & Sheehan N., 2003];
- crime risk factors analysis [Pourret et al., 2008, pp.73-84];
- credit-rating [Pourret et al., 2008, pp. 263-276].

Although the well-known discrete Bayesian networks have proved to be very useful in modelling many complex systems, most of the real-life situations require the use of hybrid networks consisting of both, continuous and discrete elements, to represent the systems most accurately.

2 Basic definitions

Introduction to continuous and hybrid Bayesian networks requires some background knowledge from the probability theory (especially referring to continuous variables) as well as from the graph theory. In the following sections some basic definitions will be provided. The presented theory can be found in [Koller D. & Friedman N., 2009, pp. 27-31, 45-61], unless stated differently.

Notation remarks:

- capital letters (e.g. X) denote random variables, lower case letters (e.g. x) stand for realisations;
- $Val(X)$ denotes all possible values of X;

- $\mathcal{X}$ designates a set of variables $X_1, ..., X_n$;
- upper-case P denotes distribution function and lower-case p stands for density function;
- $p(X, Y)$ denotes joint density of variables X and Y.
- $P(X_1, ..., X_n)$ designates joint distribution over a set of variables $\mathcal{X}$;
- BN refers to a Bayesian network.

2.1 Marginal and conditional distributions

In the Bayesian network framework two operations are of paramount importance: marginalization and conditioning on an observation. Therefore, we will now define the densities which characterize marginal and conditional distributions of random variables.

Definition 1. *(Marginal density) The marginal density of X is defined as:*

$$p(x) = \int_{-\infty}^{\infty} p(x, y) dy.$$

Hence, we can derive the marginal density by integrating the joint density $p(X, Y)$ regarding y, thus "eliminating" variable Y. It then characterizes the marginal distribution - a distribution over events that can be described using X.

A key definition for the graphical models is the notion of conditional distribution, characterized by the conditional density:

Definition 2. *(Conditional density) The conditional density of X given Y is defined as:*

$$p(x|y) = \frac{p(x, y)}{p(y)},$$

which is undefined if $p(y) = 0$.

Now, we can specify the chain rule and the Bayes' rule for the densities (the latter follows from the well-known Bayes' theorem, however, not explicitly - via a limit process):

- the chain rule: $p(x, y) = p(x)p(y \mid x)$, (for probabilities: $P(X, Y) = P(X)P(Y \mid X)$);
- the Bayes' rule: $p(x \mid y) = \dfrac{p(x)p(y \mid x)}{p(y)}$.

These notions will be referred to throughout the paper.

In the following section we will have a closer look at the graphical representation of the Bayesian network.

2.2 Bayesian network structure

A Bayesian network structure $\mathcal{G}$ is a **directed acyclic graph** (DAG) consisting of nodes, which represent random variables, and directed edges (graphical links) depicting dependencies between them. The variables are marginally independent (more information in section 2.5) if there is no edge between them. As suggested by the name, no cycles between nodes are allowed (unlike Markov graphical models).

2.3 Construction of a Bayesian network

The network structure $\mathcal{G}$ is often known beforehand, for instance from some information provided by a cooperating expert. Then, we have a simple algorithm allowing to build the network.
First, we define a topological ordering of the variables in a graph, which in this case follows from some a priori knowledge:

Definition 3. *(topological ordering) Let $\mathcal{G}$ be a graph over set of variables $\mathcal{X}$. An ordering of the nodes $X_1, ..., X_n$ is a topological ordering if $X_j \to X_i \Rightarrow j < i$.*
*Furthermore, if $X_j \to X_i$, then we call X_j a **parent** and X_i a **child**. We use $Pa_i^{\mathcal{G}}$ to denote parents of the variable X_i in the graph $\mathcal{G}$.*

Now, we can introduce the algorithm for constructing a BN:
Suppose we have some predetermined topological ordering of $\mathcal{X}$ and an empty graph $\mathcal{G}$

1. Take X_1 and set it as a root.

2. For $i = 2, ..., n$ add X_i to the network and choose its parents from $\{X_1, ..., X_{i-1}\}$ so that

$$X_i \perp (\{X_1, ..., X_{i-1}\} - Pa_i^{\mathcal{G}}) \mid Pa_i^{\mathcal{G}}.$$

   In other words: add appropriate edges between nodes and remove redundant ones.

For an example please see [Koller D. & Friedman N., 2009, pp. 79-81].

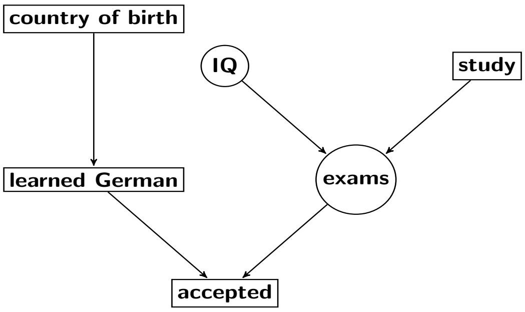
Nevertheless, in some cases we do not know how the variables depend on each other and therefore we need to construct the Bayesian network based on the data (e.g. building gene networks based on gene expression analysis). The two main network learning approaches are: constraint-based (determine dependencies through conditional independence tests) and scored-searching (identify the network maximizing a scoring function) methods. However, due to some limitations of these, hybrid learning methods have been proposed. For reference on Bayesian network learning in context of gene expression data please see [Wang M., Chen Z. & Cloutier S., 2007].

2.4 Example

Previously, the definition of a Bayesian network structure has been introduced.
Now, a "toy example" of a hybrid Bayesian network will be demonstrated and will be used throughout the next sections for explanation of the presented theory.
Suppose we are interested in the probability of being accepted to a german university (for some MA programm, e.g. management) for international students from other EU-countries.
Assumptions (taken for simplicity of the graph):

1. Admission to a MA programm requires knowledge of German, unless the applicant has outstanding academic performance (then he/she is offered a German course before the academic year).
2. The admission does not depend on other applicants in any way.
3. The average from BA studies is considered disregarding of the area of studies.

In the following (and throughout the paper) circles indicate continuous and rectangles discrete random variables.



where:
country of birth (C) = {EU-countries other than Germany } (multivariate Bernoulli distribution)
learned German (G) = {no, yes} = {0, 1} (Bernoulli distribution)
accepted (A) = {no, yes} = {0, 1} (Bernoulli distribution)

IQ (I) = intelligence of the student expressed as IQ (Gaussian distribution)
exams (E) = average from all exams taken during BA studies (Gaussian distribution)
study (S) = {no, yes} = {0, 1} = whether a person has been studying hard during BA studies (Bernoulli distribution)

2.5 Independence

A property that is crucial for constructing Bayesian networks is the independence. Due to incorporation of information about (in-)dependencies between variables, a Bayesian network can be a compact representation of a joint distribution. Typically, the number of parameters needed for a network representation is exponentially smaller.
The following section in based on [Koller D., 2013].

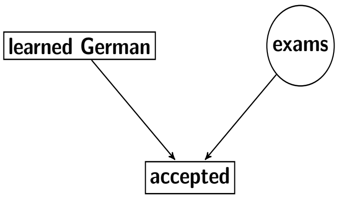
Definition 4. *(Marginal independence) We say that variables X and Y are marginally independent if $P(X, Y) = P(X)P(Y)$.*

learned German
exams

Obviously, the fact that someone has learned German does not influence his/her average from studies; similarly, exams' average has no impact on learning German. Hence, it can be suspected, that variables G and E are marginally independent.
However, additional piece of information in the network may change (in-)dependence between variables.

Definition 5. *(Conditional independence) We say that variables X and Y are conditionally independent given Z: $X \perp Y \mid Z$ if (equivalently):*

- $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$
- $P(X \mid Y, Z) = P(X \mid Z)$
- $P(Y \mid X, Z) = P(Y \mid Z)$



Introduction of an additional piece of information can both pose independence and lose it. For instance, if we introduce some information, stating whether a person has been accepted to the university or not, variables G and E are not independent anymore. Suppose the applicant has a high average from all taken exams, however he has not been accepted to the university. Then, we can conclude that the reason must have been the lack of ability of speaking German. We say, that variables G and E are conditionally dependent given A (accepted or not).

A very important notion capturing information about independencies between variables is an Independency-map (I-map).

Definition 6. *(I-map) Let $\mathcal{G}$ be any graph object associated with a set of independencies $I(\mathcal{G})$. We say that $\mathcal{G}$ is an I-map for a set of independencies I if $\mathcal{G} \subseteq I$.*

Recall the example 2.4. The joint distribution over a subset of variables $\{C, G, A, E\}$ can be decomposed (using the chain rule) as follows:

$$P(C, G, A, E) = P(E)P(C \mid E)P(G \mid C, E)P(A \mid C, E, G).$$

Nevertheless, it does not reflect the independence statements that hold in the distribution. For instance, $(A \perp C \mid G) \in \mathcal{I}(\mathcal{G})$, as the probability of being accepted to the university does not depend on the country of birth providing the information on whether the person has learned German or not. Therefore, the term $P(A \mid C, E, G)$ can be simplified to $P(A \mid E, G)$. Moreover, neither country of birth, nor learning German do not depend on the average from BA studies. Hence, the joint distribution can be decomposed as follows:

$$P(C, G, A, E) = P(E)P(C)P(G \mid C)P(A \mid E, G).$$

Remark: The definition of an I-map states that all the graph independencies must hold in the distribution, however, the graph does not necessarily represent all the independencies implied by the distribution.

To find a graph that would best, in the sense of nonredundancy, represent the joint distribution (capture all/most independencies), one would look for a minimal I-map.

Definition 7. *(minimal I-map) A graph $\mathcal{G}$ is a minimal I-map for a set of independencies $\mathcal{I}$ if it is an I-map for $\mathcal{I}$, and if the removal of even a single edge from $\mathcal{G}$ renders it not an I-map.*

Nevertheless, even such an I-map usually does not represent all of the independecies that hold in the distribution. Moreover, it is not unique and depends on the topological ordering of the variables. A minimal I-map precisely capturing all the independencies in the joint distribution is called a perfect I-map. However, it cannot be found for all possible distributions.

Due to independence assumptions, the joint distribution can be expressed as a product of individual conditional distributions:

Theorem 1. *(Factorization) Let $\mathcal{G}$ be a Bayesian network graph over the set of random variables $\mathcal{X} = X_1, ..., X_n$. We say, that a distribution P over $\mathcal{X}$ factorizes over $\mathcal{G}$ if P can be expressed as a product:*

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid Pa_i^{\mathcal{G}}),$$

where $Pa_i^{\mathcal{G}}$ denotes parents of variable X_i in graph $\mathcal{G}$.
*An individual factor $P(X_i \mid Pa_i^{\mathcal{G}})$ is called **Conditional Probability Distribution** (CPD).*

This theorem is also known as the chain rule for Bayesian networks and defines **global semantics**. Whereas **local semantics** (often referred to as Markov assumption) mean, that each node is conditionally independent of its nondescendants (variables that cannot be reached by simply following the edges) given its parents.
Furthermore, P factorizes over $\mathcal{G}$ if and only if $\mathcal{G}$ is an I-map for P. Therefore, we obtain the following definition of a Bayesian network.

Definition 8. *(Bayesian Network) A Bayesian network is a pair $\mathcal{B} = (\mathcal{G}, P)$ where P factorizes over $\mathcal{G}$, and where P is specified as a set of CPDs associated with $\mathcal{G}$'s nodes.*

The theory presented so far applies to discrete as well as continuous and hybrid networks. However, there are some important differences between these networks, which will be emphasized in the following section.

3 Continuous Bayesian networks

3.1 Introduction

Many real-life situations imply the use of continuous distributions to appropriately model a complex system. For instance, to analyse gene expression data from microarray experiments or to model systems involving temperature or height. However, the use of continuous variables causes some technical difficulties. Consider the case where both, the variable X and its parents $Y_1, ..., Y_n$ are real-valued. Then there is no representation, that could capture all possible densities, unlike in discrete case [Friedman et al., 2000]. As the domain is now real-valued, there is an unbounded number of possible parametrizations. Moreover, some technical issues arise when performing inference. Firstly, parametric family for each initial factor has to be chosen and in case of different families, situation gets very complex. Also in case of a common family chosen for all the factors, the results of even basic operations such as multiplication or marginalization may belong to a different family than initially chosen. Furthermore, as marginalization involves integration in case of continuous variables, problems related to infinite or even non-existing integrals may occur.

Nevertheless, some solutions to these problems have been developed. For instance, all the continuous variables can be discretized (see section 3.2), however on the cost of loosing some information. Another way to deal with complex multivariate continuous distributions is to use the well-known Gaussian distribution as an approximation, which (although still continuous) is easier to work with as it is closed under basic operations (see section 4).

3.2 Discretization

The most intuitive solution to deal with hybrid networks is to discretize all the continuous variables and hence, use a standard discrete model for inference. The idea behind it is to divide a continuous domain into finite set of intervals, which encapture the most of the probability mass. Then, the value of the density function is estimated over each interval, which in the simplest case means to calculate the average over each interval. As a result we obtain a table CPD, which is much more convenient to work with. Despite some limitations and loss of information, discretization remains the most common solution found in the literature to deal with continuous variables.

This section is based on [Koller D. & Friedman N., 2009, pp. 606-607].

3.2.1 Discretization procedure

Let X be a discrete or continuous child and Y its continuous parent. We replace X and Y with discrete variables A and B, respectively. Now, let $a \in Val(A)$ correspond to an interval $[y^1, y^2]$

and $b \in Val(B)$ to $[x^1, x^2]$. We define

$$P(b \mid a) = \int_{y^1}^{y^2} p(X \in [x^1, x^2] \mid Y = y) p(Y = y \mid Y \in [y^1, y^2]) dy.$$

Following this procedure allows to take into account dependencies between variables, which is crucial in Bayesian networks framework. Nevertheless, due to high computation cost of performing such a discretization a simpler approximation is often preferred:
Select a value $y^* \in [y^1, y^2]$, and estimate $P(b \mid a)$ as the total probability mass of the interval $[x^1, x^2]$ given y^*:

$$P(X \in [x^1, x^2] \mid y^*) = \int_{x^1}^{x^2} p(x \mid y^*) dx.$$

Another, slightly different approach proposed by [Kozlov A.V. & Koller D., 1997] is a dynamic discretization procedure which utilizes partition trees with nonuniform partition across all the variables to minimize the information loss. The regions of the continuous domain are assigned different importance, which results in higher accuracy and more compactness of the disrcetization. Moreover, the developed algorithm allows for self-updating, which makes it a promising tool with various applications.

3.2.2 Efficiency

Although discretization seems to be a rather good solution and is widely-used, there are some important issues that have to be taken into account. In result of discretization we can only obtain an approximation to the true distribution and hence, there might be a relatively big loss of information. Moreover, accurate approximation can only be performed on a high computation cost. The number of parameters needed to discretize distribution over c continuous variables using d discrete values for each is $O(d^c)$. Therefore, it might be advisable to use other methods, for example models involving Gaussians.

4 Gaussian models

Gaussians constitute a relatively simple class of distributions and often serve as a good approximation for other, more complex ones. Although normal distribution makes very strong assumptions (such as symmetry around mean and exponential decay away from it or linear interactions between variables [Koller D. & Friedman N., 2009, p. 247]) and these are usually not fulfilled, in many situations it is still practical to use Gaussians or their generalizations. For instance, mixtures of Gaussians are used as a solution to inference problems in hybrid Bayesian networks (see section 5.1).

The following theorems and definitions can be found in [Koller D. & Friedman N., 2009, pp. 247-253], unless stated differently.

4.1 Multivariate Gaussian distribution

Recall, that the density function of the multivariate Gaussian distribution has the form:

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp\Big[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big],$$

where $\boldsymbol{\mu}$ denotes an n-dimensional mean vector and Σ a symmetric $n \times n$ covariance matrix (with determinant $|\Sigma|$).

If $X_1, ..., X_n$ have a joint normal distribution, then X_i and X_j are independent if and only if $\Sigma_{i,j} = 0$.

For some operations (see section 4.2), however, it is more appropriate to use an alternative representation, the so-called **information form**:

$$p(\boldsymbol{x}) \propto exp\Big[-\frac{1}{2}(\boldsymbol{x}^T J \boldsymbol{x} + (J\boldsymbol{\mu})^T \boldsymbol{x}\Big],$$

where J denotes the information matrix $J = \Sigma^{-1}$.

If $X_1, ..., X_n$ have a joint normal distribution, then X_i and X_j are conditionally independent ($X_i \perp X_j \mid \mathcal{X} - \{X_i, X_j\}$) if and only if $J_{i,j} = 0$.

In both representations we usually require the covariance matrix to be positive definite and hence, invertible. Each of the forms proves to be more useful for one of the inference operations (see below) due to different computational properties.

4.2 Operations on Gaussians

As in discrete case, we can perform basic inference operations on our distribution: marginalization and conditioning.

1. **Marginalization**: obtain a marginal distribution over one of the variables from the joint distribution → use the basic parametrization

   Lemma Let $\{\boldsymbol{X}, \boldsymbol{Y}\}$ have a joint normal distribution

   $$p(\boldsymbol{X}, \boldsymbol{Y}) = \mathcal{N}\Big(\begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{pmatrix}; \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}\Big).$$

   Then the marginal distribution over $\boldsymbol{X}$ is a normal distribution $\mathcal{N}(\boldsymbol{\mu}_X; \Sigma_{XX})$. By using the basic parametrization one can easily obtain the marginal distribution by extracting it from the mean vector and covariance matrix.

2. **Conditioning**: condition the distribution on an observation $\boldsymbol{Y} = \boldsymbol{y}$ → use the information form

   $$p(\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y}) = \mathcal{N}(\boldsymbol{\mu}_X - J_{XX}^{-1} J_{XY}(\boldsymbol{y} - \boldsymbol{\mu}_Y), J_{XX}).$$

4.3 Linear Gaussians

A subclass of normal distributions that is of particular interest in terms of Bayesian networks is the class of Linear Gaussians.

Definition 9. *(Linear Gaussian model) Let X be a continuous variable with continuous parents $Y_1, ..., Y_k$. We say that X has a linear Gaussian model if there are $\beta_0, ..., \beta_k$ and σ^2 such that:*

$$P(X \mid \boldsymbol{y}) = \mathcal{N}(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{y}, \sigma^2).$$

From the definition follows that X given its parents has a **conditional linear Gaussian distribution** with the mean of X being a linear combination of the values of its parents and the variance independent of these values.

Moreover, the joint distribution of variables having a linear Gaussian distribution is a multivariate Gaussian [Lauritzen S.L. & Wermuth N., 1989]. For this model inference machinery has been well established. Therefore, even in case of a non-normal distribution, one might often assume normality for simplicity (see section 7.5 for approximation error).

4.4 Gaussian Bayesian network

If all variables in a Bayesian network are continuous and their conditional probability densities are linear Gaussians, the subsequent graphical model is called a **Gaussian Bayesian network**. Using the construction algorithm presented in 2.3 we can construct a Gaussian Bayesian network.

Theorem 2. *(Gaussian BN $\Leftarrow$ joint Gaussian) Let P be a joint Gaussian distribution over $\mathcal{X}$. Given any topological ordering $X_1, ..., X_n$ over $\mathcal{X}$, we can construct a Bayesian network graph $\mathcal{G}$ and a Bayesian network $\mathcal{B}$ such that:*

1. *$Pa_i^{\mathcal{G}} \subseteq X_1, ..., X_{i-1}$;*

2. *the CPD of X_i in $\mathcal{B}$ is a linear Gaussian of its parents;*

3. *$\mathcal{G}$ is a minimal I-map for P.*

We can easily convert a multivariate Gaussian to a linear Gaussian network (and back) and hence, these representations are equivalent. In other words, a Gaussian Bayesian network is an alternative representation of a joint Gaussian distribution.

4.5 Equivalence

The previous and the following theorems assure the equivalence between a linear Gaussian Bayesian network and a joint normal distribution. It can be shown, that the number of independent parameters in a Gaussian distribution over $X_1, ..., X_n$ and in a fully connected linear Gaussian Bayesian network is the same. Nevertheless, in general it is not possible to find a one-to-one mapping between them. The difference in parametrization results in higher compactness of a particular representation in certain situations. One can find examples where the number of parameters in the network is substantially smaller than that of the joint distribution. The opposite may also occur, a multivariate Gaussian distribution can be more compact than the network representation, for example requiring a linear number of nonzero parameters in contrast to a quadratic number in the case of the linear Gaussian Bayesian network [Koller D. & Friedman N., 2009, p. 259].

Conversion between the two representations is, however, easy and allows to use the one required for the given task.

Theorem 3. *(Gaussian BN ⇒ joint Gaussian) Let X be a linear Gaussian of its parents $Y_1, ..., Y_k$: $p(X \mid \boldsymbol{y}) = \mathcal{N}(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{y}; \sigma^2)$ Assume, that $Y_1, ..., Y_k$ are jointly Gaussian with distribution $\mathcal{N}(\boldsymbol{\mu}; \boldsymbol{\Sigma})$. Then:*

- *the distribution of X is a normal distribution $p(X) = \mathcal{N}(\mu_X; \sigma_X^2)$ where:*

$$\mu_X = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{\mu} \quad \textit{and} \quad \sigma_X^2 = \sigma^2 + \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta};$$

(μ_X being a linear function of the parents' mean and σ_X^2 obtained as a sum of the Gaussian noise parameter of X and the variance of parents weighted by the strength of the dependence)

- *the joint distribution over $\{\boldsymbol{Y}, X\}$ is a normal distribution, where:*

$$Cov[Y_i, X] = \sum_{j=0}^{k} \beta_j \boldsymbol{\Sigma_{i,j}}.$$

Consider a part of the network (2.4), where the average of exams (E) depends only on the intelligence of the person (I) (for now ignore all the other variables). Assume the following distributions:
$p(I) = \mathcal{N}(100, 15)$
$p(E \mid I) = \mathcal{N}(15 + 0.6I, 4.6) = \mathcal{N}(75, 4.6)$
Then, we can obtain the covariance matrix as follows:

$$\Sigma_{11} = 15 \qquad \Sigma_{12} = 0.6 * \Sigma_{11} = 0.6 * 15 = 9$$
$$\Sigma_{21} = \Sigma_{12} = 9 \qquad \Sigma_{22} = 4.6 + (0.6)^2 * 15 = 10$$

Hence, the marginal distribution of E is $\mathcal{N}(75, 10)$ and the joint distribution over $\{I, E\}$ is

$$\mathcal{N}\Big(\begin{pmatrix} 100 \\ 75 \end{pmatrix}; \begin{bmatrix} 15 & 9 \\ 9 & 10 \end{bmatrix}\Big)$$

5 Hybrid Bayesian networks

Networks representing most of the real-life situations involve both, discrete and continuous variables. Such models are referred to as hybrid Bayesian networks. Example applications (see [Lerner et al., 2001] for reference) include fault diagnosis (continuous: flows and pressures; discrete: failure events), visual tracking (cont.: positions of body parts; discrete: type of movement) or modelling a thermostat (cont.: temperature; discrete: heating on/off).
We consider two cases:

1. Continuous children with discrete (and continuous) parents
2. Discrete children with continuous (and discrete) parents

Remark: A network with discrete children having also continuous parents can be transformed to one with discrete children having only discrete parents by the so-called arc reversal. For more information please see [Shenoy P.P., 2006].
The following sections (5.1 and 5.2) are based on [Lerner et al., 2001].

5.1 Continuous children with discrete parents

The most common solution to the problem of modelling a system, where some of the continuous children have also discrete parents (but there are no discrete children with continuous parents) is to use linear Gaussians.

Definition 10. *(Conditional Linear Gaussian model) Let X be a continuous variable with discrete $\boldsymbol{A} = A_1, ..., A_m$ and continuous $\boldsymbol{Y} = Y_1, ..., Y_k$ parents. We define the* ***Conditional Linear Gaussian*** *(CLG) model as:*

$$p(X \mid \boldsymbol{a}, \boldsymbol{y}) = \mathcal{N}(w_{\boldsymbol{a},0} + \sum_{i=1}^{k} w_{\boldsymbol{a},i} y_i; \sigma_a^2),$$

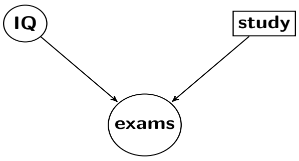
where w are coefficients.

As in the linear Gaussian model, the mean is a linear function of the continuous parents. The covariance is still independent of continuous parents, but can be influenced by the assignment to discrete ones. Please note, that the above definition is redundant as one of the linear combinations can be expressed as (1 - all the other combinations).

What makes this model so widely used is its property of defining a conditional Gaussian joint distribution: for each assignment to the discrete variables (parents), we get a separate linear Gaussian model for the child. Thus, the distribution over continuous variables (children) is a multivariate Gaussian. Although it is certainly a restricted model, it proved to be useful in various situations.

A Bayesian network with all discrete variables having only discrete parents and continuous variables having a CLG CPD is called a **Conditional Linear Gaussian network**. Such networks are also being referred to as mixtures of Gaussians Bayesian networks and are widely used in inference as approximation to other distributions [Shenoy P.P., 2006].

Recall the example 2.4. Consider the continuous variable *exams*(E) with both, continuous (IQ(I)) and discrete (*study*(S)) parents.



The conditional distribution of *exams*, given its parents is: $\mathcal{N}(85 - 20s, 10 + 2s)$ as:
$p(I) = \mathcal{N}(100, 15)$
$p(E \mid I, S = 1) = \mathcal{N}(25 + 0.6I, 10) = \mathcal{N}(85, 10)$
$p(E \mid Q, S = 0) = \mathcal{N}(-5 + 0.7I, 12) = \mathcal{N}(65, 12)$

5.2 Discrete children with continuous parents

Examples, where a continuous quantity influences a discrete variable are ubiquitous. For instance, consider a situation of a person having fever. Depending on the temperature (continuous) a doctor might draw conclusions on whether to provide the patient with medication or not. We can assume, that we have some threshold which determines the change in discrete values. In example 2.4 we could consider such a threshold in the average from exams, which determines the probability of being accepted to the university. For instance 75 points, then:
$f(E) \geq 75 \Rightarrow P(A = 1)$ likely to be 1
$f(E) < 75 \Rightarrow P(A = 1)$ likely to be 0

Such and similar situations are rather problematic in terms of modelling and require a special class of CPDs, namely **augmented Conditional Linear Gaussians**. They were first defined by [Lerner et al., 2001] as a generalization of a standard softmax function. Although there are many possible representations, the softmax or logit function proved to be the most useful as it solves the problem of dicontinuity arised by introducing the threshold.

Definition 11. *(Augmented CLG) Let A be a discrete variable with possible values* $a_1, ..., a_m$ *and let* $\boldsymbol{Y} = Y_1, ..., Y_k$ *denote its continuous parents. We define the CPD in augmented Conditional*

Linear Gaussian model as:

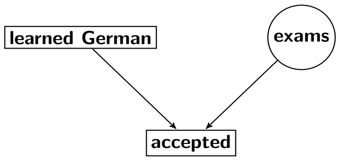
$$P(A = a_i \mid y_1, ..., y_k) = \frac{exp(w_0^i + \sum_{l=1}^{k} w_l^i y_l)}{\sum_{j=1}^{m} exp(w_0^j + \sum_{l=1}^{k} w_l^j y_l)},$$

where w are coefficients.

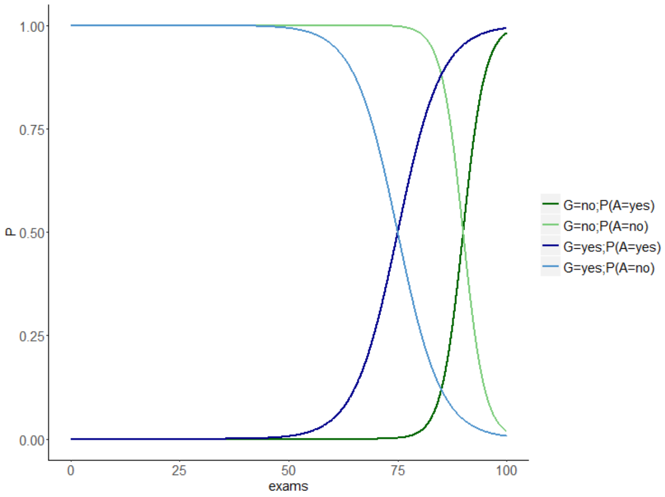
The above definition is redundant, like in CLG case. Thus, dividing the above numerator and denominator by $exp(w_0^i + \sum_{l=1}^{k} w_l^i y_l)$ leads to a CPD with m-1 parameters.
Furthermore, we also allow the case where the discrete child has both, continuous and discrete parents and we model it by defining a softmax function for each combination of discrete parents (like in case of CLGs). In case of a binary child, the model simplifies to a sigmoid CPD:

$$P(A = a_1 \mid y_1, ..., y_k,) = \frac{1}{1 + exp(w_0 + \sum_{l=1}^{k} w_l y_l)}.$$

In our example 2.4, this way of modelling is very useful while considering the following part of the Bayesian network:



We have, namely, a discrete variable (*accepted* (A)) with both, discrete (*learned German*(G)) as well as continuous (*exams*(E)) parents.



In the above plot (created with ggplot2, RStudio) we can see how the probability of being accepted (not being accepted - lighter colour) changes with the average from all the exams in two cases: the applicant has learned German and not. In the first case the threshold in continuous values (*exams*) can be estimated as 75 points, whereas in the second one - as 90 points. Hence, it is clear that learning German sufficiently shifts the threshold, so that a person with no knowledge of German would have less chances of being accepted having the same grades as the one learning German. The green plot is steeper, which indicates that small changes in the exams average have bigger influence on the probability of acceptance. In both cases, the probability saturates towards 1 as the exams average reaches 100 points.

5.3 Inference

Various algorithms have been developed to handle hybrid Bayesian networks. As previously stated (section 3), inference in both, purely continuous as well as in hybrid networks, poses some technical difficulties. Nevertheless, a solution for mixtures of Gaussians (MoG) proposed by [Lauritzen S.L. & Jensen, 2001] has been well established. In case of augmented CLG networks and non-Gaussian distributions, approximation by MoG have been suggested. This can be achieved for instance by numerical integration, dynamic discretization or sampling methods (e.g. EM-algorithm); the latter ones being approximate methods. For reference please see [Shenoy P.P., 2006]. For details on inference in hybrid Bayesian networks please see [Koller D. & Friedman N., 2009, pp. 605-647].

6 Exponential family

In previous sections the focus was put on single distributions. Now, going a step further, we will consider families of distributions with special attention paid to the most common and very important one, namely the exponential family. Among distributions belonging to this family are Gaussian-, Poisson-, exponential-, geometric-, Gamma distributions and many more. However, not all of the commonly used distributions are in this family, including the Student t-distribution. Why is this family so important in the context of Bayesian networks? In certain situations it is crucial to find an approximation to the true distribution, which is relatively easy to perform within the same family. Furthermore, exponential families have some properties, that make them relatively easily computable. On top of that, due to robustness they often arise in terms of optimization (see section 7 for entropy) and several other Bayes' procedures.

Examples of exponential family models are ubiquitous and include: Ising model (statistical physics), Metric Labelling and Potts model, Gaussian Markov Random Field, Latent Dirichlet Allocation and Models with Hard-Constraints (communication theory). For reference and more information please see [Wainwright M. & Jordan M., 2008, pp. 41-51].

This section serves as foundation for network learning and inference (in both, discrete and continuous cases) and is based on [Koller D. & Friedman N., 2009, pp. 261-269].

6.1 Definition

Definition 12. *(Exponential family) An exponential family $\mathcal{P}$ over a set of variables $\mathcal{X}$ is specified by:*

1. *A sufficient statistics function τ from assignments to $\mathcal{X}$ to R^K;*
2. *A parameter space that is a convex set $\Theta \subseteq R^M$ of legal parameters;*
3. *A natural parameter function t from R^M to R^K;*
4. *An auxiliary measure A over $\mathcal{X}$.*

Each vector of parameters $\boldsymbol{\theta} \subseteq \Theta$ specifies a distribution P_θ in the family as

$$P_\theta(\xi) = \frac{1}{Z(\theta)} A(\xi) exp\{\langle t(\boldsymbol{\theta}), \tau(\xi)\rangle\}, \tag{1}$$

where $\langle t(\boldsymbol{\theta}), \tau(\xi)\rangle$ is the inner product of the vectors $t(\theta)$ and $\tau(\xi)$, and

$$Z(\boldsymbol{\theta}) = \sum_\xi A(\xi) exp\{\langle t(\boldsymbol{\theta}), \tau(\xi)\rangle\}$$

is the partition function of $\mathcal{P}$, which must be finite. The parametric family $\mathcal{P}$ is defined as:

$$\mathcal{P} = P_\theta : \boldsymbol{\theta} \in \Theta.$$

The definition might seem rather complex, but as we will see on an example, it provides quite an elegant way of presenting the distribution function.

For instance, for a univariate Gaussian distribution define:

$$\tau(x) = \langle x, x^2\rangle; \quad t(\mu, \sigma^2) = \langle \frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\rangle; \quad A(\xi) = 1.$$

Then

$$Z(\mu, \sigma^2) = \sqrt{2\pi}\sigma exp\left\{\frac{\mu^2}{2\sigma^2}\right\}$$

and

$$P(x) = \frac{1}{Z(\mu, \sigma^2)} exp\{\langle t(\theta), \tau(X)\} = \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

We would like our exponential family to be invertible, so we want the function t to be invertible over the set of parameters. It facilitates many operations, or even makes them possible in the first place. An exponential family is inveritble if and only if it is nonredundant. In other words, if parameters are different, then the distributions differ as well ($\theta \neq \theta^{'} \Rightarrow P_\theta \neq P_{\theta'}$) and thus, the family is invertible. Moreover, we demand the parameter space to be convex and open in R^M. Convexity is a characteristic that facilitates optimization tasks. If additionally the parameter space has the following form, called **natural parameter space**:

$$\Theta = \left\{\boldsymbol{\theta} \in \mathcal{R}^K : \int exp\{\langle \boldsymbol{\theta}, \tau(\xi)\rangle\} d\xi < \infty\right\}$$

and the function t is the identity, we obtain parameters (called natural parameters) of the same dimension as the data representation. Such a family is called a **linear exponential family** and provides a concise representation of the distribution. Some nonlinear distributions may be reparametrized to obtain a linear exponential family.

6.2 Factors

For inference operations it is often most convenient to represent the distribution in a factor form:

Definition 13. *An unnormalized factor in an exponential family is:* $\phi_{\boldsymbol{\theta}}(\xi) = A(\xi) exp\{\langle t(\boldsymbol{\theta}), \tau(\xi)\rangle\}$.

A composition (product) of such factors also belongs to the exponential family. This remark is sufficient to conclude that a product of exponential CPDs (factorized distribution) stays within the exponential family.

6.3 Bayesian networks with exponential CPDs

In the light of the above, a Bayesian network, in which CPDs of all the variables belong to an exponential family, defines an exponential family. However, it is not necessarily a linear family of distributions.

An important subtlety of such exponential networks is that we can construct a Bayesian network only then, when all the CPDs are locally normalized (which is easy to perform). A counter-example showing that a global normalization constant is not sufficient can be found in [Koller D. & Friedman N., 2009, p. 268]. As the normalization is performed locally, a Bayesian network can be seen as an exponential family with the partition function (global term) equal to 1.

7 Entropy

In previous sections we assumed, that the true distribution of the data is known, however it is often not the case. To introduce the tools able to handle such a situation, we have to introduce the notion of entropy.

Entropy was primarily defined by Rudolf Clausius (1865) in the field of thermodynamics as a measure of degree of disorder in a system. It has also applications in other areas of knowledge and is sometimes referred to as the amount of "noise".

This section is based on [Koller D. & Friedman N., 2009, pp. 269-273, 1137-1142], unless stated differently.

7.1 Statistical entropy

The statistical notion of entropy (entropy of a distribution) was defined by [Shannon C.E., 1948] and is also called the Shannon's Measure of Uncertainty.

Definition 14. *(Statistical entropy)*
The entropy of a distribution P over X is defined as

$$\boldsymbol{H}_P(X) = \boldsymbol{E}_P\Big[log\frac{1}{P(X)}\Big] = -\boldsymbol{E}_P[logP(X)].$$

Properties:

1. $\bigvee_P \ \boldsymbol{H}_P(X) \geq 0$;

2. $\exists_{x_0} \ P(X = x_0) = 1 \Rightarrow \boldsymbol{H}_P(X) = 0$;

3. $\boldsymbol{H}_P$ is the largest when probabilities for all values are equal;

4. $0 \leq \boldsymbol{H}_P(X) \leq log|Val(X)|$.

The entropy of a distribution provides us with information on how the distribution mass is spread:
low entropy ⇒ distribution mass is on a few instances
high entropy ⇒ distribution mass is widely spread
In other words, the larger the entropy is, the more uniform is the distribution.

7.1.1 Differential entropy

Entropy of a continuous variable is also called differential entropy. Although the concept is similar to the entropy of a discrete variable, there are some differences worth mentioning. This section is based on [Cover T.M. & Thomas J.A., 2006, pp. 243-255].

Definition 15. *(Differential entropy) The differential entropy $\boldsymbol{H}(X)$ of a continuous random variable X with density $f(x)$ is defined as*

$$\boldsymbol{H}(X) = -\int_S f(x) log f(x) dx,$$

where S is the support set of the random variable.

The definition holds only if the above integral exists and is well defined.
Properties:

1. $\boldsymbol{H}(X + c) = \boldsymbol{H}(X)$;
2. $\boldsymbol{H}(aX) = \boldsymbol{H}(X) + log|a|$ ($\boldsymbol{H}$ is not invariant under the change of variables);
3. $\boldsymbol{H}$ lacks nonnegativity (example: uniform distribution).

The above mentioned nuances should be taken into account while computing entropy in the case of a continuous variable.

7.2 Maximum Entropy Principle

Consider the following situation: we have an empirical dataset without much knowledge about the true distribution. We have, however, information about some dependencies (e.g. from previous studies). Then, to model the distribution of the data, from all distributions we choose the one, that matches our additional piece of information (referred to as constraints) and has the highest entropy (**Maximum Entropy Principle**). In this way we assure against the worst case, where the distribution is the least informative. Moreover, we do not force any specific characteristics that may not hold in the true distribution and thus, we avoid making false assumptions about the data or introducing bias. Please note, that this method is not limited to the continuous case.

The Maximum Entropy approach has been especially successful in medical diagnostics, where we often encounter insufficient statistical information. Even though collaborating with experts usually allows for the correct determination of the network structure, quantifying conditional probabilities remains a challenge. The main advantage of this method is the requirement for local computations only (depends on the constraints), which makes it efficient even for large scale models. For more information see [Wiegernick W. & Heskes T., 2001].
To obtain a distribution with maximum entropy given some constraints, the most common method is to use the Lagrange multipliers. An example, which proves the following theorem can be found in [Conrad K., n.d., p.25].

Theorem 4. *For the class of distributions with known mean and variance, the normal distribution is the one with maximum entropy.*

7.3 Entropy of a distribution in an exponential family

To obtain the entropy in many cases we can make use of the exponential form of the distribution. In case of an exponential family, computing entropy seems to be a relatively easy task, as we use the expectation of the sufficient statistics under P_θ rather than assignments to $\mathcal{X}$.

Theorem 5. *(Entropy within an exponential family) Let P_θ be a distribution in an exponential family defined by the functions τ and t. Then,*

$$\boldsymbol{H}_P(X) = lnZ(\boldsymbol{\theta}) - \langle \boldsymbol{E}_{P_\theta}[\tau(\mathcal{X})], t(\boldsymbol{\theta})\rangle.$$

Example: Recall the univariate Gaussian distribution (section 6.1). The entropy can be calculated as follows:

$$\boldsymbol{H}_P = ln(\sqrt{2\pi}\sigma exp\left\{\frac{\mu^2}{2\sigma^2}\right\}) - \left(\boldsymbol{E}_P[X]\frac{\mu}{\sigma^2} + \boldsymbol{E}_P[X^2](-\frac{1}{2\sigma^2})\right) =$$

$$= \frac{1}{2}ln(2\pi\sigma^2) + \frac{\mu^2}{2\sigma^2} - \left(\boldsymbol{E}_P[X]\frac{\mu}{\sigma^2} + (\boldsymbol{Var}_P[X] + (\boldsymbol{E}_P[X])^2)(-\frac{1}{2\sigma^2})\right) =$$

$$= \frac{1}{2}ln(2\pi\sigma^2) + \frac{\mu^2}{2\sigma^2} - \frac{\mu^2}{\sigma^2} + \frac{\sigma^2}{2\sigma^2} + \frac{\mu^2}{2\sigma^2} = \frac{1}{2}ln(2\pi\sigma^2) + \frac{1}{2} = \frac{1}{2}ln(2\pi e\sigma^2).$$

We can see that in case of Gaussians, the bigger the variance, the larger the entropy is.

What is interesting, solving a problem of finding a distribution with maximum entropy while posting linear constraints leads to an exponential family. Furthermore, if the constraint is a given mean, then the distribution with maximum entropy has necessarily the form of an exponential family [Wainwright M. & Jordan M., 2008, p.63].

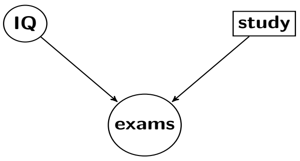
7.4 Entropy of a Bayesian network

Computing the overall entropy of a Bayesian network requires local considerations. In such, we can use the exponential form of the CPDs to facilitate computations. However, we also need to obtain the weighting term - the marginal distribution of the parents, which may depend on some other entries in the network.

Theorem 6. *(Entropy of a Bayesian network) If* $P(\mathcal{X}) = \prod_i P(X_i \mid Pa_i^{\mathcal{G}})$ *is a distribution consistent with a Bayesian network* $\mathcal{G}$*, then*

$$\boldsymbol{H}_P(\mathcal{X}) = \sum_i \boldsymbol{H}_P(X_i \mid Pa_i^{\mathcal{G}}) = \sum_i \sum_{pa_i^{\mathcal{G}}} P(pa_i^{\mathcal{G}}) \boldsymbol{H}_P(X_i \mid pa_i^{\mathcal{G}}).$$

Hence, we can compute the entropy of a Bayesian network simply by summing the conditional entropies of all distributions composing the joint distribution consistent with the network. Therefore, we can conclude that if all conditional entropies are low/ high, then the overall entropy is low/high, respectively. Please note, that the above theorem holds for discrete parents only. A generalisation to the case of hybrid networks is possible, but requires further considerations and treating each kind of variables (with discrete, with continuous and with mixed parents) differently. Most importantly, for continuous parents an integral should be incorporated into the formula. For a simplified demonstration, recall the example 2.4. Consider a part of the network: the continuous variable *exams* with both, continuous (*IQ*) and discrete (*study*) parents.



Variables *IQ* (I) and *study* (S) have no parents, we can compute the entropy directly from the marginal distributions. Moreover, I is normally distributed and S follows the Bernoulli distribution, which both belong to the exponential family. Therefore, we can use the previously shown formula from Theorem 5 (section 7.3) to obtain the following results:

$\boldsymbol{H}_P(I) = \frac{1}{2}ln(2\pi e\sigma^2)$;

$\boldsymbol{H}_P(S) = -\alpha ln(\alpha) - (1-\alpha)ln(1-\alpha)$ where $\alpha = P(S=1)$ probability of studying a lot.

The variable *exams* has two parents: I and S. Thus, the conditional entropy:

$\boldsymbol{H}_P(E \mid I, S) \overset{*}{=} \int_{Val(I)} \sum_{Val(S)} p(i)P(S=s)\boldsymbol{H}_P(E \mid i, S=s)di =$

$= \int_{Val(I)}(p(i)P(S=1)\boldsymbol{H}_P(E \mid i, S=1) + p(i)P(S=0)\boldsymbol{H}_P(E \mid i, S=0))di =$

$\overset{**}{=} \int_{Val(I)} p(i)[\alpha\frac{1}{2}ln(2\pi e\sigma_1^2) + (1-\alpha)\frac{1}{2}ln(2\pi e\sigma_0^2)]di \overset{***}{=} \alpha\frac{1}{2}ln(2\pi e\sigma_1^2) + (1-\alpha)\frac{1}{2}ln(2\pi e\sigma_0^2)$.

$*$ I and S are marginally independent.

$**$ As conditional distribution of E given its parents is normal, we can use the formula for entropy of a distribution within the exponential family.

$***$ The expression in [ ] does not depend on i and $\int_{Val(I)} p(i)di = 1$ as density.

Hence, the overall entropy of this part of the network is:

$\boldsymbol{H}_P(I, S, E) = \frac{1}{2}ln(2\pi e\sigma^2) - \alpha ln(\alpha) - (1-\alpha)ln(1-\alpha) + \alpha\frac{1}{2}ln(2\pi e\sigma_1^2) + (1-\alpha)\frac{1}{2}ln(2\pi e\sigma_0^2)$.

7.5 Relative entropy

Consider the following situation: we have a dataset from which we obtain the empirical distribution. As we do not know the true distribution from which the data were generated, we try to find the best approximation. Thus, we are interested in how close we are from the true distribution. In other words, we would like to know how much information we loose due to approximation. The relative entropy, also known as the Kullback-Leibler divergence (or KL-distance), provides an answer to this question.

7.5.1 General case

Definition 16. *(Relative entropy) Let Q and P be two distributions over random variables $X_1, ..., X_n$. The relative entropy of Q and P is:*

$$\boldsymbol{D}(Q(X_1, ..., X_n) \parallel P(X_1, ..., X_n)) = \boldsymbol{E}_Q\left[log\frac{Q(X_1, ..., X_n)}{P(X_1, ..., X_n)}\right],$$

where we set log(0) = 0.

Properties:

1. $\bigvee_{P,Q} \boldsymbol{D}(Q \parallel P) \geq 0$;
2. $\boldsymbol{D}(Q \parallel P) = 0 \Leftrightarrow P = Q$ (almost everywhere);
3. $\bigvee_{P \neq Q} \boldsymbol{D}(Q \parallel P) \neq D(P \parallel Q)$;
4. $\boldsymbol{D}(Q \parallel P)$ small $\Rightarrow P$ close to $Q \Rightarrow$ small loss of information.

Another example when we would use the relative entropy measure, is when the true distribution is somewhat hard to work with due to complex representation or does not possess qualities we might need to perform inference. Then, we approximate this distribution with another one (in the continuous case usually with Gaussian) and need to control how far, in the sense of relative entropy, we are from the true distribution.

7.5.2 Relative entropy in Bayesian networks

If we would like to measure the relative entropy between distributions, from which at least one is consistent with a Bayesian network, then we could use the following theorem:

Theorem 7. *(Relative entropy in a Bayesian network) If P is a distribution over $\mathcal{X}$ consistent with a Bayesian network $\mathcal{G}$, then*

$$\boldsymbol{D}(Q \parallel P) = -\boldsymbol{H}_Q(\mathcal{X}) - \sum_i \sum_{pa_i^{\mathcal{G}}} Q(pa_i^{\mathcal{G}}) \boldsymbol{E}_{Q(X_i, pa_i^{\mathcal{G}})}[lnP(X_i \mid pa_i^{\mathcal{G}})].$$

If Q is also consistent with $\mathcal{G}$, then

$$\boldsymbol{D}(Q \parallel P) = \sum_i \sum_{pa_i^{\mathcal{G}}} Q(pa_i^{\mathcal{G}}) \boldsymbol{D}(Q(X_i, pa_i^{\mathcal{G}})) \parallel P(X_i \mid pa_i^{\mathcal{G}})).$$

Although the formula might seem complex, the terms to be summed decompose into weights (depending on the joint distribution of Q) and either the expected value (the first result) or the relative entropy between conditional distributions (the second result).

8 Projections

Recall the previously described situations, where we want to approximate the true distribution (P), because it is either too complex or unknown. We would like to use some distribution (Q) with desired properties and keep the relative entropy small. A solution to this problem comes

with the projection operation, where we project (by analogy to geometry) a distribution onto a family of distributions ($\mathcal{Q}$), e.g. the exponential family, a family of factored distributions, by minimizing the KL-distance between them.
As this operation explicitly depends on the relative entropy, which is not symmetrical, we distinguish two types of projections: information- and moment- projection.
This section is based on [Koller D. & Friedman N., 2009, pp. 273-283].

8.1 I-projections

Definition 17. *(I-projection) The I-projection (information projection) of a distribution P onto family $\mathcal{Q}$ is defined as*

$$Q^I = arg \min_{Q \in \mathcal{Q}} \boldsymbol{D}(Q \parallel P) = arg \min_{Q \in \mathcal{Q}} (-\boldsymbol{H}_Q(X) + \boldsymbol{E}_Q[-lnP(X)]).$$

This type of projections is usually used when the given model is hard to work with and we would like to answer some probability queries. It assigns high density to regions where P is large and low to those where P is small. Nevertheless, to avoid the situation where all the probability mass is put on a region most probable according to P, there is a penalty for low entropy. Although some simplification of computation is possible, obtaining an I-projection usually requires some advanced mathematical machinery.

8.2 M-projections

Definition 18. *(M-projection) The M-projection (moment projection) of distribution P onto family $\mathcal{Q}$ is defined as*

$$Q^M = arg \min_{Q \in \mathcal{Q}} \boldsymbol{D}(P \parallel Q) = arg \min_{Q \in \mathcal{Q}} (-\boldsymbol{H}_P(X) + \boldsymbol{E}_P[-lnQ(X)]).$$

Finding an M-projection is useful especially while learning a graphical model from data. In other words, when we are looking for a distribution within a particular family matching given dataset. This type of projections attempts to match the main mass of P by assigning high density to the regions that are probable according to the true distribution. Furthermore, there is a high penalty for assigning low density in these regions. Thus, the obtained distribution is characterized by a relatively high variance (due to high density in all regions in support of P).

Sometimes using the exponential form of the distribution may simplify the computation, as in this case we only need to compare expected sufficient statistics of both distributions.

Theorem 8. *(M-projection onto the exponential family) Let P be a distribution over $\mathcal{X}$, and $\mathcal{Q}$ a set of distributions within the exponential family defined by the functions $\tau(\xi)$ and $t(\boldsymbol{\theta})$. If there is a set of parameters $\boldsymbol{\theta}$ such that $\boldsymbol{E}_{Q_\theta}[\tau(\mathcal{X})] = \boldsymbol{E}_P[\tau(\mathcal{X})]$, then the M-projection of P is Q_θ.*

Moreover, in the exponential family sufficient statistics are often moments and then the M-projection can be found just by searching the family for a distribution matching the moments of P. For details on the proceeding and explanation of the additional machinery please see [Koller D. & Friedman N., 2009, p. 278].

In the Bayesian network framework, the task of finding an M-projection (in the sense of projecting a distribution onto a Bayesian network) simplifies to factorization of the distribution P.

Theorem 9. *(M-projection onto Bayesian network) Let P be a distribution over $X_1, ..., X_n$, and let $\mathcal{G}$ be a Bayesian network structure. Then the M-projection Q^M is:*

$$Q^M(X_1, ..., X_n) = \prod_i P(X_i \mid Pa_i^{\mathcal{G}}),$$

which follows from the theorem on relative entropy in Bayesian networks (see section 7.5.2).

9 Summary

Discrete Bayesian networks have proved to be useful in many situations, however continuous and hybrid networks have been receiving more and more attention as they model many systems that include continuous quantities (e.g. gene expression) more accurately. Despite the many advantages of these graphical models, introducing continuous variables to the network causes some technical problems in terms of both, representation and inference. The most intuitive and still most common method to circumvent these issues, discretization, faces some important limitations due to the loss of information. Therefore, another solution has been proposed, in which the normal distribution is used as an approximation for other continuous distributions, as Gaussians are closed under the basic inference operations. For discrete-continuous cases the Conditional Linear Gaussian model have been proposed and for continuous-discrete ones an augmented version, using the softmax function. In many cases finding an appropriate distribution as approximation can be facilitated by the use of the exponential form. The loss of information caused by approximating can be measured by means of relative entropy and thus, minimizing the KL-distance between distributions provides a considerably good approximation. Especially when the true distribution is unknown or has a complex structure we may want to project the empirical distribution onto a family of distributions (e.g. Gaussians). As emphasized in this seminar paper, more attention should be paid to continuous variables as they often arise in real-life situations. Although many tools have been developed to incorporate such information to a Bayesian network, still much remains to be done.

References

[1] Conrad K., n.d., *Probability distributions and maximum entropy* Expository papers: Analysis, University of Connecticut CT, USA, URL http://www.math.uconn.edu/ kconrad/blurbs/analysis/entropypost.pdf [access: 03.01.2016], p. 25

[2] Cover T.M. & Thomas J.A., 2006, *Elements of Information Theory*, 2nd edition, A Wiley-Interscience publication, ISBN: 10 0-471-24195-4, John Wiley & Sons, Inc., pp. 243-255

[3] Frey B.J., 1998, *Graphical Models for Machine Learning and Digital Communication*, ISBN: 978-0-262-06202-2, The MIT Press, p. 58

[4] Friedman N., Linial M., Nachman I. & Pe'er D., 2000, *Using Bayesian Networks to Analyze Expression Data*, Journal of Computational Biology, Vol. 7, no. 3/4, pp.127-135

[5] Koller D. & Friedman N., 2009, *Probabilistic Graphical Models. Principles and Techniques* within Dietterich T. (eds.) *Adaptive computation and machine learning*, ISBN: 978-0-262-01319-2, The MIT Press, pp. 27-31, 45-61, 79-81, 247-253, 261-283, 605-647, 1137-1142

[6] Koller D., 2013, on-line course "Probabilistic Graphical Models", URL https://class.coursera.org/pgm/lecture [access: 03.01.2016]

[7] Kozlov A.V. & Koller D., 1997, *Nonuniform dynamic discretization in hybrid networks*, Proceedings of the 13th Conference Annual Conference on Uncertainty in Artifical Intelligence, pp. 314-325

[8] Lauritzen S.L. & Jensen F., 2001, *Stable local computation with conditional Gaussian distributions*, Statistics and Computing, Vol. 11, no. 2, pp. 191-203

[9] Lauritzen S.L. & Sheehan N., 2003, *Graphical Model for Genetic Analyses*, Statistical Science, Vol. 18, No. 4, pp. 489-514

[10] Lauritzen, S.L. & Wermuth, N., 1989, *Graphical Models for Associations between Variables, some of which are Qualitative and some Quantitative*, The Annals of Statistics, Vol. 17, no. 1, pp. 31-57

[11] Lerner U., Segal E. & Koller D., 2001, *Exact Inference in Networks with Discrete Children of Continuous Parents*, Proceedings of the 17th Conference Annual Conference on Uncertainty in Artificial Intelligence, pp. 319-328

[12] Nikovski D., 2000, *Constructing Bayesian networks for medical diagnosis from incomplete and partially correct statistics* EEE Transactions on Knowledge and Data Engineering, Vol. 12, No. 4, pp. 509-516

[13] Pourret O., Naim P. & Marcot B., 2008, *Bayesian networks. A practical guide to applications*, within Scott M.(eds.), Senn S. (eds.) & Barnett V. (eds.) *Statistics in practice*, ISBN: 978-0-470-06030-8, John Wiley & Sons Ltd, pp. 73-84, 263-276

[14] RStudio Team, 2015, *RStudio: Integrated Development Environment for R*, RStudio, Inc., URL http://www.rstudio.com/

[15] Shannon C.E., 1948, *A Mathematical Theory of Communication.*, The Bell System Technical Journal, Vol. 27, Issue 3, pp. 379-423

[16] Shenoy P.P., 2006, *Inference in Hybrid Bayesian Networks Using Mixtures of Gaussians*, Proceedings of the 22nd Conference Annual Conference on Uncertainty in Artificial Intelligence, pp.428-436

[17] Wainwright M. & Jordan M., 2008, *Graphical Models, Exponential Families, and Variational Inference*, Foundations and Trends in Machine Learning, Vol. 7, Issue 1-2, pp. 41-51

[18] Wang M., Chen Z. & Cloutier S., 2007, *A hybrid Bayesian network learning method for constructing gene networks*, Computational Biology and Chemistry, Vol. 31, Issue 5-6, pp. 361-372

[19] Wickham H., 2009, *ggplot2: Elegant Graphics for Data Analysis*, ISBN: 978-0-387-98140-6, Springer-Verlag New York

[20] Wiegerinck W. and Heskes T., 2001, *Probability Assessment with Maximum Entropy in Bayesian Networks*, Proceedings of the 33rd Symposium on the Interface, Computating Science and Statistics, pp. 183-191