

Bayesian network and information theory

Ait-taleb Nabil *

Abstract

In this paper, we will expose the BIC score expressed as a function of the Bayesian network's entropy. We will then use this BIC score to learn a Bayesian network from an example of data frame.

*Corresponding author: nabiltravail1982@gmail.com

1 Introduction

In this paper, we will begin by defining the notions of joint and conditional differential entropy that we find in the field of information theory. We will expose the Gaussian multivariate probability and compute its joint differential entropy and conditional differential entropy. We will then give the definition of the Bayesian network and how show to compute its entropy.

We will also expose the entropy of two particular Bayesian networks: one with a chain rule and one without edges. These two particular Bayesian networks will allow us to understand, using an inequality, the range of values in which the entropy of a Bayesian network varies.

We will present in a section the likelihood function and the BIC score expressed from a Bayesian network's entropy. We will then describe an algorithm allowing us to obtain the most likely optimal Bayesian network from a data frame using the BIC score expressed from the Bayesian network's entropy. We will briefly explain what to do to infer data from the optimal Bayesian network.

This paper will end with the application of the learning algorithm on a concrete example of a data frame.

2 Differential entropy attributed to random vectors

2.1 Differential entropy for a random vector

Definition: Given a random vector \vec{x} , defined on set \mathbb{X} of size n , with a multidimensional probability density function (pdf) $p_X(\vec{x})$, we define the differential entropy $h(X)$ as:

$$h(X) = - \int_{\mathbb{X}} p_X(\vec{x}) \ln p_X(\vec{x}) \overrightarrow{d\vec{x}}$$

2.2 Joint differential entropy of two random vectors

Definition: Given two concatenated random vectors (\vec{x}_1, \vec{x}_2) , defined on the sets \mathbb{X}_1 and \mathbb{X}_2 of size n and m respectively, with a multidimensional probability density function(pdf) $p_{X_1 X_2}(\vec{x}_1, \vec{x}_2)$, we define the joint differential entropy $h(X_1, X_2)$ as:

$$h(X_1, X_2) = - \int_{\mathbb{X}_1} \int_{\mathbb{X}_2} p_{X_1 X_2}(\vec{x}_1, \vec{x}_2) \ln p_{X_1 X_2}(\vec{x}_1, \vec{x}_2) \overrightarrow{dx_1 dx_2}$$

2.3 Conditional differential entropy of a random vector given a random vector

Definition: Given two concatenated random vectors (\vec{x}_1, \vec{x}_2) , defined on the sets \mathbb{X}_1 and \mathbb{X}_2 of size n and m respectively, with a multidimensional probability density function(pdf) $p_{X_1 X_2}(\vec{x}_1, \vec{x}_2)$, we define the conditional differential entropy $h(X_1|X_2)$ as:

$$h(X_1|X_2) = - \int_{\mathbb{X}_1} \int_{\mathbb{X}_2} p_{X_1 X_2}(\vec{x}_1, \vec{x}_2) \ln p_{X_1|X_2}(\vec{x}_1, \vec{x}_2) \overrightarrow{dx_1 dx_2}$$

where we have:

$$p_{X_1|X_2}(\vec{x}_1, \vec{x}_2) = \frac{p_{X_1 X_2}(\vec{x}_1, \vec{x}_2)}{p_{X_2}(\vec{x}_2)}$$

$$P_{X_2}(\vec{x}_2) = \int_{\mathbb{X}_1} p_{X_1 X_2}(\vec{x}_1, \vec{x}_2) \overrightarrow{dx_1}$$

3 Multivariate Gaussian distribution and information theory

3.1 Joint and conditional gaussian multidimensional probability

Consider a partitioned random vector $\vec{x} = (\vec{x}_1, \vec{x}_2)$ of size $n = k_1 + k_2$, where k_1 and k_2 are the size of vectors \vec{x}_1 and \vec{x}_2 respectively, with a multivariate Gaussian distribution $P_X(\vec{x})$ with a mean vector $\vec{\mu}_X$ and covariance matrix K_{X^2} :

$$P_X(\vec{x}) = \mathcal{N}(\mu_X, K_{X^2}) = (2\pi)^{-\frac{n}{2}} |K_{X^2}|^{-\frac{1}{2}} \exp\left\{-\frac{(\vec{x} - \vec{\mu}_X)^T K_{X^2}^{-1} (\vec{x} - \vec{\mu}_X)}{2}\right\}$$

The purpose of this section is to expose the following different probabilities:

1. $P_X(\vec{x}) = P_{X_1, X_2}(\vec{x}_1, \vec{x}_2)$
2. $P_{X_2}(\vec{x}_2)$
3. $P_{X_1|X_2}(\vec{x}_1, \vec{x}_2)$

For this, we must start first from the block matrix multiplication of the covariance matrix K and the precision matrix $W = K^{-1}$ and prove the following relation:

$$W_{X_2^2} = K_{X_2^2}^{-1} + W_{X_2 X_1} \cdot W_{X_1^2}^{-1} \cdot W_{X_1 X_2} \cdot$$

Indeed:

$$K_{X^2} W_{X^2} = \begin{pmatrix} K_{X_1^2} W_{X_1^2} + K_{X_1 X_2} W_{X_2 X_1} & K_{X_1^2} W_{X_1 X_2} + K_{X_1 X_2} W_{X_2^2} \\ K_{X_2 X_1} W_{X_1^2} + K_{X_2^2} W_{X_2 X_1} & K_{X_2 X_1} W_{X_1 X_2} + K_{X_2^2} W_{X_2^2} \end{pmatrix} = \begin{pmatrix} I_{k_1, k_1} & 0 \\ 0 & I_{k_2, k_2} \end{pmatrix}$$

$$K_{X_2 X_1} W_{X_1^2} + K_{X_2^2} W_{X_2 X_1} = 0$$

$$K_{X_2^2}^{-1} K_{X_2 X_1} W_{X_1^2} + W_{X_2 X_1} = 0$$

$$K_{X_2^2}^{-1} K_{X_2 X_1} = -W_{X_2 X_1} W_{X_1^2}^{-1}$$

$$K_{X_2^2} W_{X_2^2} + K_{X_2 X_1} W_{X_1 X_2} = I_{k_2, k_2}$$

$$W_{X_2^2} = K_{X_2^2}^{-1} - K_{X_2^2}^{-1} \cdot K_{X_2 X_1} \cdot W_{X_1 X_2}$$

Finally, we obtain:

$$W_{X_2^2} = K_{X_2^2}^{-1} + W_{X_2 X_1} \cdot W_{X_1^2}^{-1} \cdot W_{X_1 X_2}$$

Now, we will develop the Mahalanobis distance:

$$\begin{aligned}
& (\bar{x} - \bar{\mu}_X)' W_{X^2} (\bar{x} - \bar{\mu}_X) \\
&= (\bar{x}_1 - \bar{\mu}_{\bar{X}_1}, \bar{x}_2 - \bar{\mu}_{\bar{X}_2}) \begin{pmatrix} W_{X_1^2} & W_{X_1 X_2} \\ W_{X_2 X_1} & W_{X_2^2} \end{pmatrix} \begin{pmatrix} \bar{x}_1 - \bar{\mu}_{\bar{X}_1} \\ \bar{x}_2 - \bar{\mu}_{\bar{X}_2} \end{pmatrix} \\
&= (\bar{x}_1 - \bar{\mu}_{\bar{X}_1})' W_{X_1^2} (\bar{x}_1 - \bar{\mu}_{\bar{X}_1}) + (\bar{x}_1 - \bar{\mu}_{\bar{X}_1})' W_{X_1 X_2} (\bar{x}_2 - \bar{\mu}_{\bar{X}_2}) + (\bar{x}_2 - \bar{\mu}_{\bar{X}_2})' W_{X_2 X_1} (\bar{x}_1 - \bar{\mu}_{\bar{X}_1}) \\
&+ (\bar{x}_2 - \bar{\mu}_{\bar{X}_2})' W_{X_2^2} (\bar{x}_2 - \bar{\mu}_{\bar{X}_2})
\end{aligned}$$

Using the relation: $W_{X_2^2} = K_{X_2^2}^{-1} + W_{X_2 X_1} \cdot W_{X_1^2}^{-1} \cdot W_{X_1 X_2}$, we obtain:

$$\begin{aligned}
&= (\bar{x}_1 - \bar{\mu}_{\bar{X}_1})' W_{X_1^2} (\bar{x}_1 - \bar{\mu}_{\bar{X}_1}) + (\bar{x}_1 - \bar{\mu}_{\bar{X}_1})' W_{X_1 X_2} (\bar{x}_2 - \bar{\mu}_{\bar{X}_2}) + (\bar{x}_2 - \bar{\mu}_{\bar{X}_2})' W_{X_2 X_1} (\bar{x}_1 - \bar{\mu}_{\bar{X}_1}) \\
&+ (\bar{x}_2 - \bar{\mu}_{\bar{X}_2})' W_{X_2 X_1} W_{X_1^2}^{-1} W_{X_1 X_2} (\bar{x}_2 - \bar{\mu}_{\bar{X}_2}) + (\bar{x}_2 - \bar{\mu}_{\bar{X}_2})' K_{X_2^2}^{-1} (\bar{x}_2 - \bar{\mu}_{\bar{X}_2}) \\
&= [(\bar{x}_1 - \bar{\mu}_{\bar{X}_1}) + W_{X_1^2}^{-1} W_{X_1 X_2} (\bar{x}_2 - \bar{\mu}_{\bar{X}_2})]' [W_{X_1^2} (\bar{x}_1 - \bar{\mu}_{\bar{X}_1}) + W_{X_1 X_2} (\bar{x}_2 - \bar{\mu}_{\bar{X}_2})] \\
&+ (\bar{x}_2 - \bar{\mu}_{\bar{X}_2})' K_{X_2^2}^{-1} (\bar{x}_2 - \bar{\mu}_{\bar{X}_2}) \\
&= [(\bar{x}_1 - \bar{\mu}_{\bar{X}_1}) + W_{X_1^2}^{-1} W_{X_1 X_2} (\bar{x}_2 - \bar{\mu}_{\bar{X}_2})]' W_{X_1^2} \cdot [(\bar{x}_1 - \bar{\mu}_{\bar{X}_1}) + W_{X_1^2}^{-1} W_{X_1 X_2} (\bar{x}_2 - \bar{\mu}_{\bar{X}_2})] \\
&+ (\bar{x}_2 - \bar{\mu}_{\bar{X}_2})' K_{X_2^2}^{-1} (\bar{x}_2 - \bar{\mu}_{\bar{X}_2})
\end{aligned}$$

We put:

$$Q_1 = (\bar{x}_1 - \nu_{X_1/X_2})' (K_{X_1^2} - K_{X_1 X_2} K_{X_2^2}^{-1} K_{X_2 X_1})^{-1} (\bar{x}_1 - \nu_{X_1/X_2})$$

$$\nu_{\bar{X}_1|X_2} = \bar{\mu}_{\bar{X}_1} + K_{X_1 X_2} K_{X_2^2}^{-1} (\bar{x}_2 - \bar{\mu}_{\bar{X}_2})$$

$$Q_2 = (\bar{x}_2 - \bar{\mu}_{\bar{X}_2})' K_{X_2^2}^{-1} (\bar{x}_2 - \bar{\mu}_{\bar{X}_2})$$

We then obtain the equalities as follows:

$$(\bar{x} - \bar{\mu}_X)' K_{X^2}^{-1} (\bar{x} - \bar{\mu}_X) = Q_1 + Q_2$$

$$P_X(\bar{x}) = (2\pi)^{-\frac{n}{2}} |K_{X^2}|^{-\frac{1}{2}} \exp\left\{-\frac{Q_1+Q_2}{2}\right\} = (2\pi)^{-\frac{n}{2}} |K_{X^2}|^{-\frac{1}{2}} \exp\left\{-(\bar{x} - \bar{\mu}_X)' K_{X^2}^{-1} (\bar{x} - \bar{\mu}_X)\right\}$$

$$P_{X_2}(\bar{x}_2) = (2\pi)^{-\frac{k_2}{2}} |K_{X_2^2}|^{-\frac{1}{2}} \exp\left\{-\frac{Q_2}{2}\right\} = (2\pi)^{-\frac{k_2}{2}} |K_{X_2^2}| \exp\left\{-\frac{(\bar{x}_2 - \bar{\mu}_{\bar{X}_2})' K_{X_2^2}^{-1} (\bar{x}_2 - \bar{\mu}_{\bar{X}_2})}{2}\right\}$$

Using the relation $\frac{P_{X_1 X_2}(\bar{x}_1, \bar{x}_2)}{P_{X_2}(\bar{x}_2)}$:

$$P_{X_1|X_2}(\bar{x}_1, \bar{x}_2) = (2\pi)^{-\frac{k_1}{2}} \left(\frac{|K_{(X_1, X_2)}|^2}{|K_{X_2^2}|}\right)^{-\frac{1}{2}} \exp\left\{-\frac{Q_1}{2}\right\} = (2\pi)^{-\frac{k_1}{2}} \left(\frac{|K_{(X_1, X_2)}|^2}{|K_{X_2^2}|}\right)^{-\frac{1}{2}} \exp\left\{-\frac{(\bar{x}_1 - \nu_{\bar{X}_1|X_2})' (K_{X_1^2} - K_{X_1 X_2} K_{X_2^2}^{-1} K_{X_2 X_1})^{-1} (\bar{x}_1 - \nu_{\bar{X}_1|X_2})}{2}\right\}$$

If we use the Schur's complement $K_{X_1^2|X_2} = K_{X_1^2} - K_{X_1 X_2} K_{X_2^2}^{-1} K_{X_2 X_1}$

we can express the conditional probability $P_{X_1|X_2}(\bar{x}_1, \bar{x}_2)$ as follows:

$$\begin{aligned}
& P_{X_1|X_2}(\bar{x}_1, \bar{x}_2) \\
&= (2\pi)^{-\frac{k_1}{2}} \left(\frac{|K_{(X_1, X_2)}|^2}{|K_{X_2^2}|}\right)^{-\frac{1}{2}} \exp\left\{-\frac{(\bar{x}_1 - \nu_{\bar{X}_1|X_2})' K_{X_1^2|X_2}^{-1} (\bar{x}_1 - \nu_{\bar{X}_1|X_2})}{2}\right\} = (2\pi)^{-\frac{k_1}{2}} |K_{X_1^2|X_2}|^{-\frac{1}{2}} \exp\left\{-\frac{(\bar{x}_1 - \nu_{\bar{X}_1|X_2})' K_{X_1^2|X_2}^{-1} (\bar{x}_1 - \nu_{\bar{X}_1|X_2})}{2}\right\}
\end{aligned}$$

3.2 Differential entropy of a Gaussian random vector

Theorem: Given random vector $\vec{x} = (x_1, x_2, \dots, x_n)$ with a multivariate Gaussian distribution:

$$P_X(\vec{x}) = \mathcal{N}(\mu_X, K_{X^2}) = (2\pi)^{-\frac{n}{2}} |K_{X^2}|^{-\frac{1}{2}} \exp\left\{-\frac{(\vec{x} - \vec{\mu}_X)^t K_{X^2}^{-1} (\vec{x} - \vec{\mu}_X)}{2}\right\}$$

with a mean vector μ_X and a covariance matrix K_{X^2} then the differential entropy is equal to:

$$h(X) = \frac{1}{2} \ln(2\pi e)^n |K_{X^2}|$$

Proof:

$h(X)$

$$\begin{aligned} &= - \int_{-\infty}^{+\infty} p_X(\vec{x}) \ln\{p_X(\vec{x})\} \vec{d}\vec{x} \\ &= - \int_{-\infty}^{+\infty} p_X(\vec{x}) \left[-\frac{1}{2} (\vec{x} - \mu_X)^t K_{X^2}^{-1} (\vec{x} - \mu_X) - \ln(\sqrt{2\pi})^n |K_{X^2}|^{\frac{1}{2}} \right] \vec{d}\vec{x} \\ &= \frac{1}{2} E_X \left[\sum_{ij} (\vec{x}_i - \mu_{X_i})^t (K_{X^2}^{-1})_{ij} (\vec{x}_j - \mu_{X_j}) \right] + \frac{1}{2} \ln(2\pi)^n |K_{X^2}| \\ &= \frac{1}{2} E_X \left[\sum_{ij} (\vec{x}_i - \mu_{X_i})^t (\vec{x}_j - \mu_{X_j}) (K_{X^2}^{-1})_{ij} \right] + \frac{1}{2} \ln(2\pi)^n |K_{X^2}| \\ &= \frac{1}{2} \sum_{ij} E_X [(\vec{x}_j - \mu_{X_j})^t (\vec{x}_i - \mu_{X_i})] (K_{X^2}^{-1})_{ij} + \frac{1}{2} \ln(2\pi)^n |K_{X^2}| \\ &= \frac{1}{2} \sum_{ij} [(K_{X^2})_{ji} (K_{X^2}^{-1})_{ij}] + \frac{1}{2} \ln(2\pi)^n |K_{X^2}| \\ &= \frac{1}{2} \sum_j [(K_{X^2})_{jj} (K_{X^2}^{-1})_{jj}] + \frac{1}{2} \ln(2\pi)^n |K_{X^2}| \\ &= \frac{1}{2} \sum_j I_{jj} + \frac{1}{2} \ln(2\pi)^n |K_{X^2}| \\ &= \frac{n}{2} + \frac{1}{2} \ln(2\pi)^n |K_{X^2}| \\ &= \frac{1}{2} \ln(2\pi e)^n |K_{X^2}| \end{aligned}$$

3.3 Conditional differential entropy of two Gaussian random vectors

Theorem: Given two concatenated Gaussian random vectors $\vec{x} = (\vec{x}_1, \vec{x}_2)$, of size k_1 and k_2 respectively, with a multivariate Gaussian distribution:

$$P_X(\vec{x}) = \mathcal{N}(\mu_X, K_{X^2}) = (2\pi)^{-\frac{n}{2}} |K_{X^2}|^{-\frac{1}{2}} \exp\left\{-\frac{(\vec{x} - \vec{\mu}_X)' K_{X^2}^{-1} (\vec{x} - \vec{\mu}_X)}{2}\right\}$$

with a mean vector μ_X and a covariance matrix K_{X^2} .

In this case, the conditional differential entropy $h(X_1|X_2)$ is equal to :

$$h(X_1|X_2) = \frac{1}{2} \ln(2\pi e)^{k_1} |K_{X_1^2|X_2}|$$

Proof:

$$h(X_1|X_2) = - \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p_{X_1, X_2}(\vec{x}_1, \vec{x}_2) \ln\{p_{X_1|X_2}(\vec{x}_1, \vec{x}_2)\} \overrightarrow{dx_1} \overrightarrow{dx_2}$$

We know the conditional probability $P_{X_1|X_2}$ can be expressed as follows:

$$P_{X_1|X_2}(\vec{x}_1, \vec{x}_2) = (2\pi)^{-\frac{k_1}{2}} |K_{X_1^2|X_2}|^{-\frac{1}{2}} \exp\left\{-\frac{(\vec{x}_1 - \nu_{X_1|X_2})' K_{X_1^2|X_2}^{-1} (\vec{x}_1 - \nu_{X_1|X_2})}{2}\right\}$$

So we can write:

$$\begin{aligned} h(X_1|X_2) &= - \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p_{X_1, X_2}(\vec{x}_1, \vec{x}_2) \left[-\frac{1}{2} (\vec{x}_1 - \nu_{X_1|X_2})' K_{X_1^2|X_2}^{-1} (\vec{x}_1 - \nu_{X_1|X_2}) - \ln(\sqrt{2\pi})^{k_1} |K_{X_1^2|X_2}|^{\frac{1}{2}} \right] \overrightarrow{dx_1} \overrightarrow{dx_2} \\ &= \frac{1}{2} E_{X_1, X_2} \left[\sum_{ij} \{(\vec{x}_1)_i - \nu_{(X_1|X_2)_i}\}' (K_{X_1^2|X_2}^{-1})_{ij} \{(\vec{x}_1)_j - \nu_{(X_1|X_2)_j}\} \right] + \frac{1}{2} \ln(2\pi)^{k_1} |K_{X_1^2|X_2}| \\ &= \frac{1}{2} E_{X_1, X_2} \left[\sum_{ij} \{(\vec{x}_1)_i - \nu_{(X_1|X_2)_i}\}' \{(\vec{x}_1)_j - \nu_{(X_1|X_2)_j}\} (K_{X_1^2|X_2}^{-1})_{ij} \right] + \frac{1}{2} \ln(2\pi)^{k_1} |K_{X_1^2|X_2}| \\ &= \frac{1}{2} \sum_{ij} E_{X_1, X_2} \left[\{(\vec{x}_1)_j - \nu_{(X_1|X_2)_j}\}' \{(\vec{x}_1)_i - \nu_{(X_1|X_2)_i}\} \right] (K_{X_1^2|X_2}^{-1})_{ij} + \frac{1}{2} \ln(2\pi)^{k_1} |K_{X_1^2|X_2}| \\ &= \frac{1}{2} \sum_{ij} (K_{X_1^2|X_2})_{ji} (K_{X_1^2|X_2}^{-1})_{ij} + \frac{1}{2} \ln(2\pi)^{k_1} |K_{X_1^2|X_2}| \\ &= \frac{1}{2} \sum_j (K_{X_1^2|X_2})_{jj} (K_{X_1^2|X_2}^{-1})_{jj} + \frac{1}{2} \ln(2\pi)^{k_1} |K_{X_1^2|X_2}| \\ &= \frac{1}{2} \sum_j I_{jj} + \frac{1}{2} \ln(2\pi)^{k_1} |K_{X_1^2|X_2}| \\ &= \frac{k_1}{2} + \frac{1}{2} \ln(2\pi)^{k_1} |K_{X_1^2|X_2}| \\ &= \frac{1}{2} \ln(2\pi e)^{k_1} |K_{X_1^2|X_2}| \end{aligned}$$

3.4 Inequalities theorem on the conditional differential entropies gaussian vectors

Theorem: Given a partitioned Gaussian random vector $\vec{x} = (\vec{x}_1, \vec{x}_2, \vec{x}_3)$, of size k_1 , k_2 and $k_3 = 1$ respectively, with the multivariate Gaussian distribution $\mathcal{N}(\mu_X, K_X)$ then we can write the following inequalities:

$$h(X_3|X_1, X_2) \leq h(X_3|X_1) \leq h(X_3)$$

Proof:

For this, we must start first from the block matrix multiplication of the covariance matrix $K_{(X_1, X_2)^2}$ and the precision matrix $W_{(X_1, X_2)^2} = K_{(X_1, X_2)^2}^{-1}$ and prove the following relation:

$$W_{X_1^2} = K_{X_1^2}^{-1} + W_{X_1 X_2} \cdot W_{X_2^2}^{-1} \cdot W_{X_2 X_1}$$

$$K_{X^2} W_{X^2} = \begin{pmatrix} K_{X_1^2} W_{X_1^2} + K_{X_1 X_2} W_{X_2 X_1} & K_{X_1^2} W_{X_1 X_2} + K_{X_1 X_2} W_{X_2^2} \\ K_{X_2 X_1} W_{X_1^2} + K_{X_2^2} W_{X_2 X_1} & K_{X_2 X_1} W_{X_1 X_2} + K_{X_2^2} W_{X_2^2} \end{pmatrix} = \begin{pmatrix} I_{k_1, k_1} & 0 \\ 0 & I_{k_2, k_2} \end{pmatrix}$$

$$K_{X_1 X_2} \cdot W_{X_2^2} + K_{X_1^2} \cdot W_{X_1 X_2} = 0$$

$$K_{X_1^2}^{-1} \cdot K_{X_1 X_2} \cdot W_{X_2^2} + W_{X_1 X_2} = 0$$

$$K_{X_1^2}^{-1} \cdot K_{X_1 X_2} = -W_{X_1 X_2} \cdot W_{X_2^2}^{-1}$$

$$K_{X_2^2} \cdot W_{X_2^2} + K_{X_1 X_2} \cdot W_{X_2 X_1} = I_{k_1 k_1}$$

$$W_{X_1^2} = K_{X_1^2}^{-1} - K_{X_1^2}^{-1} \cdot K_{X_1 X_2} \cdot W_{X_2 X_1}$$

$$W_{X_2^2} = K_{X_2^2}^{-1} + W_{X_1 X_2} \cdot W_{X_2^2}^{-1} \cdot W_{X_2 X_1}$$

We must develop the following quadratic form for $n = k_1 + k_2 + k_3 = k_1 + k_2 + 1$:

$$\begin{aligned} & (K_{X_3, X_1}, K_{X_3, X_2}) \cdot K_{(X_1, X_2)^2}^{-1} \cdot \begin{pmatrix} K_{X_1, X_3} \\ K_{X_2, X_3} \end{pmatrix} \\ &= (K_{X_3, X_1}, K_{X_3, X_2}) \cdot W_{(X_1, X_2)^2} \cdot \begin{pmatrix} K_{X_1, X_3} \\ K_{X_2, X_3} \end{pmatrix} \\ &= (K_{X_3, X_1}, K_{X_3, X_2}) \begin{pmatrix} W_{X_1^2} & W_{X_1 X_2} \\ W_{X_2 X_1} & W_{X_2^2} \end{pmatrix} \begin{pmatrix} K_{X_1, X_3} \\ K_{X_2, X_3} \end{pmatrix} \\ &= (K_{X_3, X_1})^t W_{X_1^2} (K_{X_1, X_3}) + (K_{X_3, X_1}) W_{X_1 X_2} (K_{X_2, X_3}) + (K_{X_3, X_2}) W_{X_2 X_1} (K_{X_1, X_3}) \\ &+ (K_{X_3, X_2}) \cdot W_{X_2^2} (K_{X_2, X_3}) \end{aligned}$$

Using the relation: $W_{X_2^2} = K_{X_2^2}^{-1} + W_{X_1 X_2} \cdot W_{X_2^2}^{-1} \cdot W_{X_2 X_1}$:

$$= (K_{X_3, X_1}) \cdot K_{X_1^2}^{-1} \cdot (K_{X_1, X_3}) + (K_{X_3, X_1}) \cdot W_{X_1 X_2} \cdot W_{X_2^2}^{-1} \cdot W_{X_2 X_1} \cdot (K_{X_1, X_3})$$

$$\begin{aligned}
& + (K_{X_3 X_1}) \cdot W_{X_1 X_2} \cdot (K_{X_2 X_3}) + (K_{X_3 X_2}) \cdot W_{X_2 X_1} \cdot (K_{X_1 X_3}) \\
& + (K_{X_3 X_2}) \cdot W_{X_2^2} \cdot (K_{X_2 X_3}) \\
& = [(K_{X_2 X_3}) + W_{X_2^2}^{-1} \cdot W_{X_2 X_1} \cdot (K_{X_1 X_3})]^t \cdot [W_{X_2^2} \cdot (K_{X_2 X_3}) + W_{X_2 X_1} \cdot (K_{X_1 X_3})] \\
& + (K_{X_3 X_1}) \cdot K_{X_1^2}^{-1} \cdot (K_{X_1 X_3}) \\
& = (K_{X_3 X_1}) \cdot K_{X_1^2}^{-1} \cdot (K_{X_1 X_3}) \\
& + [(K_{X_2 X_3}) + W_{X_2^2}^{-1} \cdot W_{X_2 X_1} \cdot (K_{X_1 X_3})]^t \cdot W_{X_2^2} \cdot [(K_{X_2 X_3}) + W_{X_2^2}^{-1} \cdot W_{X_2 X_1} \cdot (K_{X_1 X_3})]
\end{aligned}$$

However both following quadratics forms are equivalents :

$$(K_{X_3 X_1}, K_{X_3 X_2}) \begin{pmatrix} W_{X_1^2} & W_{X_1 X_2} \\ W_{X_2 X_1} & W_{X_2^2} \end{pmatrix} \begin{pmatrix} K_{X_1 X_3} \\ K_{X_2 X_3} \end{pmatrix} = (K_{X_3 X_2}, K_{X_3 X_1}) \begin{pmatrix} W_{X_2^2} & W_{X_2 X_1} \\ W_{X_1 X_2} & W_{X_1^2} \end{pmatrix} \begin{pmatrix} K_{X_2 X_3} \\ K_{X_1 X_3} \end{pmatrix}$$

and are positive semidefinite if and only if:

$$W_{X_1^2} \geq 0, W_{X_2^2} - W_{X_2 X_1} \cdot W_{X_1^2}^{-1} \cdot W_{X_1 X_2} \geq 0$$

but yet

$$W_{X_2^2} \geq 0, W_{X_1^2} - W_{X_1 X_2} \cdot W_{X_2^2}^{-1} \cdot W_{X_2 X_1} \geq 0$$

As $W_{X_2^2} \geq 0$, we can write the inequalities:

$$(K_{X_3 X_1}, K_{X_3 X_2}) \cdot K_{(X_1 X_2)^2}^{-1} \cdot \begin{pmatrix} K_{X_1 X_3} \\ K_{X_2 X_3} \end{pmatrix} \geq (K_{X_3 X_1}) \cdot K_{X_1^2}^{-1} \cdot (K_{X_1 X_3}) \geq 0$$

If we use $K_{X_3^2}$, we can write:

$$K_{X_3^2} - (K_{X_3 X_1}, K_{X_3 X_2}) \cdot K_{(X_1 X_2)^2}^{-1} \cdot \begin{pmatrix} K_{X_1 X_3} \\ K_{X_2 X_3} \end{pmatrix} \leq K_{X_3^2} - (K_{X_3 X_1}) \cdot K_{X_1^2}^{-1} \cdot (K_{X_1 X_3}) \leq K_{X_3^2}$$

$$K_{X_3^2|X_1 X_2} \leq K_{X_3^2|X_1} \leq K_{X_3^2}$$

$$\frac{1}{2} \ln |K_{X_3^2|X_1 X_2}| + \frac{1}{2} \ln(2\pi e)^n \leq \frac{1}{2} \ln |K_{X_3^2|X_1}| + \frac{1}{2} \ln(2\pi e)^n \leq \frac{1}{2} \ln |K_{X_3^2}| + \frac{1}{2} \ln(2\pi e)^n$$

Finally, we have the relation:

$$h(X_3|X_1, X_2) \leq h(X_3|X_1) \leq h(X_3)$$

4 Bayesian network

4.1 Bayesian network definition

Definition A Bayesian network \mathcal{B} is a directed acyclic graph having n nodes which verify the following proprieties:

- For each node, we attribute n random variables included in a random vector $\vec{x} = (x_1, x_2, \dots, x_n)$
- For each node, we attribute a conditional probability $P_{X_j|Pa(X_j)}(x_j, Pa(x_j))$ corresponding to the probabilities of child random variables x_j given the parents random variables $Pa(x_j)$ on the graph related to Bayesian network.
- The Bayesian network verify the following factorization rule:

$$p_X(\vec{x}|\mathcal{B}) = p_X(x_1, x_2, \dots, x_n|\mathcal{B}) = \prod_{X_j \in X} p_{X_j|Pa(X_j)}(x_j, Pa(x_j))$$

where we have:

$$X = \{X_j\}_{j=1,2,\dots,n}$$

$$p_{X_j|Pa(X_j)}(x_j, Pa(x_j)) = \frac{p_{X_j, Pa(X_j)}(x_j, Pa(x_j))}{p_{Pa(X_j)}(Pa(x_j))}$$

In what follows, we will consider **Gaussian** random vectors $\vec{x} = (x_1, x_2, \dots, x_n)$

4.2 Differential entropy of a Bayesian network

If \mathcal{B} is a Bayesian network to which we attribute a Gaussian random vector (X_1, X_2, \dots, X_n) , where the variables $X_{j=1,\dots,n}$ belong to the the set of Gaussian random variables X , we can compute the differential entropy of this network as follows:

$$\begin{aligned} h(X|\mathcal{B}) &= -E_X[\ln p_X(\vec{x}|\mathcal{B})] \\ &= -E_X[\ln \prod_{X_j \in X} p_{X_j|Pa(X_j)}(x_j, Pa(x_j))] \\ &= -E_X[\sum_{X_j \in X} \ln p_{X_j|Pa(X_j)}(x_j, Pa(x_j))] \\ &= \sum_{X_j \in X} -E_X[\ln p_{X_j|Pa(X_j)}(x_j, Pa(x_j))] \\ &= \sum_{X_j \in X} h(X_j|Pa(X_j)) \\ &= \frac{1}{2} \sum_{X_j \in X} \ln(2\pi e) K_{X_j^2|Pa(X_j)} \\ &= \frac{1}{2} \sum_{X_j \in X} \ln(K_{X_j^2|Pa(X_j)}) + \frac{1}{2} \ln(2\pi e)^n \end{aligned}$$

4.3 Entropy for a Bayesian network with a chain rule

If \mathcal{B}^C is a Bayesian network with a chain rule to which we attribute a Gaussian random vector (X_1, X_2, \dots, X_n) , where the variables $X_{j=1,2,\dots,n}$ belong to the set of Gaussian random variables X , we can compute the differential entropy $h(X|\mathcal{B}^C)$ of Bayesian network \mathcal{B}^C as follows:

$$h(X|\mathcal{B}^C) = h(X_1, X_2, \dots, X_n) = h(X_1) + \sum_{j=2}^n h(X_j|X_1, \dots, X_{j-1}) = \frac{1}{2} \ln |K_{X^2}| + \frac{1}{2} \ln(2\pi e)^n$$

Note: This relation is invariant by the permutation on the nodes: we choose any order on the order of the nodes in the chain and we will obtain the same result for the entropy value.

4.4 Entropy for a Bayesian network without edges

If \mathcal{B}^R is a Bayesian network without edges to which we attribute a Gaussian random vector (X_1, X_2, \dots, X_n) , where the variables $X_{j=1,2,\dots,n}$ belong to the set of Gaussian random variables X , we can compute the differential entropy $h(X|\mathcal{B}^R)$ of Bayesian network \mathcal{B}^R as follows:

$$h(X|\mathcal{B}^R) = \sum_{j=1}^n h(X_j) = \sum_{j=1}^n \ln(K_{X_{jj}^2}) + \frac{1}{2} \ln(2\pi e)^n$$

4.5 Lower bound and upper bound of a Bayesian network's differential entropy

Theorem Given a Gaussian random vector $\vec{x} = (x_1, x_2, \dots, x_n)$ with a multivariate Gaussian distribution $\mathcal{N}(\mu_X, K_{X^2})$ assigned to the nodes of a Bayesian network \mathcal{B} , then the entropy of this Bayesian network can be bounded as follows:

$$h(X|\mathcal{B}^C) \leq h(X|\mathcal{B}) \leq h(X|\mathcal{B}^R)$$

$$\frac{1}{2} \ln |K_{X^2}| + \frac{1}{2} \ln(2\pi e)^n \leq h(X|\mathcal{B}) \leq \frac{1}{2} \sum_{i=1}^n \ln(K_{X_i^2}) + \frac{1}{2} \ln(2\pi e)^n$$

$$h(X_1, X_2, \dots, X_n) \leq \sum_{X_j \in X} h(X_j | Pa(X_j)) \leq \sum_{X_j \in X} h(X_j)$$

where we have X_j , the following inequalities for each node :

$$h(X_j | Pa(X_j), Pa^C(X_j)) \leq h(X_j | Pa(X_j)) \leq h(X_j)$$

where $Pa^C(X_j) = X \setminus \{Pa(X_j)\}$

The lower bound is computed from the closure of a graph related to the Bayesian network.

The upper bound is computed by removing the set of edges on the graph related to the Bayesian network.

Proof:

Let's consider a Bayesian network's factorization performed in topological order $\mathcal{O}(X)$:

$$p_X(\vec{x}|\mathcal{B}) = p_X(x_1, x_2, \dots, x_n|\mathcal{B}) = \prod_{X_j \in \mathcal{O}(X)} p_{X_j | Pa(X_j)}(x_j, Pa(x^j))$$

We attribute a Bayesian network's factorization which graph is the closure of the graph related to the initial Bayesian network (this is a chain in the topological order):

$$p_X(\vec{x}|\mathcal{B}^C) = p_X(x_1, x_2, \dots, x_n|\mathcal{B}^C) = \prod_{X_j \in \mathcal{O}(X)} p_{X_j | Pa(X_j), Pa^C(X_j)}(x_j, Pa(x^j), Pa^C(X_j))$$

and a Bayesian network's factorization computed by removing the set of edges of graph related to the initial Bayesian network:

$$p_X(\vec{x}|\mathcal{B}^R) = p_X(x_1, x_2, \dots, x_n|\mathcal{B}^R) = \prod_{X_j \in \mathcal{O}(X)} p_{X_j}(x_j)$$

The entropies of the three bayesian networks can be computed as follows:

$$h(X|\mathcal{B}) = \sum_{x_j \in \mathcal{O}(X)} h(X_j|Pa(X_j))$$

$$h(X|\mathcal{B}^C) = \sum_{x_j \in \mathcal{O}(X)} h(X_j|Pa(X_j), Pa^C(X_j))$$

$$h(X|\mathcal{B}^R) = \sum_{x_j \in \mathcal{O}(X)} h(X_j)$$

However we proved: $h(X_3|X_1, X_2) \leq h(X_3|X_1) \leq h(X_3)$

We can write for each node X_j the relation:

$$h(X_j|Pa(X_j), Pa^C(X_j)) \leq h(X_j|Pa(X_j)) \leq h(X_j)$$

we obtain the lower and upper boundaries for the Bayesian network's entropy:

$$\sum_{X_j \in \mathcal{O}(X)} h(X_j|Pa(X_j), Pa^C(X_j)) \leq \sum_{X_j \in \mathcal{O}(X)} h(X_j|Pa(X_j)) \leq \sum_{X_j \in \mathcal{O}(X)} h(X_j)$$

The product of the conditional variances $K_{X_j^2|Pa(X_j), Pa^C(X_j)}$ and therefore of the Schur's complements give us the determinant of matrix K_{X^2} .

Therefore, the following results are computed for the lower boundary:

$$\begin{aligned} & h(X|\mathcal{B}^C) \\ &= h(X_1, X_2, \dots, X_n) \\ &= \sum_{x_j \in \mathcal{O}(X)} h(X_j|Pa(X_j), Pa^C(X_j)) \\ &= \frac{1}{2} \ln \left(\prod_{x_j \in \mathcal{O}(X)} K_{X_j^2|Pa(X_j), Pa^C(X_j)} \right) + \frac{1}{2} \ln(2\pi \cdot e)^n \\ &= \frac{1}{2} \ln(|K_{X^2}|) + \frac{1}{2} \ln(2\pi \cdot e)^n \end{aligned}$$

and for the upper boundary :

$$\begin{aligned} & h(X|\mathcal{B}^R) \\ &= \sum_{x_j \in \mathcal{O}(X)} h(X_j) \\ &= \frac{1}{2} \sum_{x_j \in \mathcal{O}(X)} \ln(K_{X_j^2}) + \frac{1}{2} \ln(2\pi e)^n \\ &= \frac{1}{2} \sum_{i=1}^n \ln(K_{X_i^2}) + \frac{1}{2} \ln(2\pi e)^n \end{aligned}$$

Finally, we obtain:

$$\frac{1}{2} \ln(|K_{X^2}|) + \frac{1}{2} \ln(2\pi.e)^n \leq \sum_{X_j \in \mathcal{O}(X)} h(X_j | Pa(X_j)) \leq \frac{1}{2} \sum_{i=1}^n \ln(K_{X_i^2}) + \frac{1}{2} \ln(2\pi e)^n$$

Note that we can use this theorem to prove that the determinant of a symmetric positive semidefinite matrix is always less than or equal to the product of the diagonal elements of this matrix. This inequality is called *Hadamard's inequality*.(see appendix)

5 The likelihood function, the BIC score and Bayesian network's entropy

Given a Gaussian random vector $X = \{x_{j=1,2,\dots,n}\}$ and a data frame D of size $N \times n$.

As we have the relation:

$$\mu_{(X_j|Pa(X_j))} = \mu_{X_j} + K_{(X_j,Pa(X_j))} \cdot K_{Pa^2(X_j)}^{-1} (Pa(X_j) - \mu_{X_j})$$

We can put:

$$\beta_{X_j} = \mu_{X_j} - K_{(X_j,Pa(X_j))} \cdot K_{Pa^2(X_j)}^{-1} \cdot \mu_{Pa(X_j)}$$

$$\beta_{(X_j,Pa(X_j))} = K_{(X_j,Pa(X_j))} \cdot K_{Pa^2(X_j)}^{-1}$$

For each current node X_j , we can write the multivariate Gaussian distribution as follows:

$$P_{X_j|Pa(X_j)}(x_j, Pa(x_j)) = (2\pi)^{-\frac{1}{2}} K_{X_j^2|Pa(X_j)}^{-\frac{1}{2}} \exp\left\{-\frac{(x_j - \beta_{(X_j,Pa(X_j))}) \cdot Pa(x_j) - \beta_{X_j})^2}{2 \cdot K_{X_j^2|Pa(X_j)}}\right\}$$

For N points of a data frame D , we then compute the likelihood function:

$$\begin{aligned} L(D|\beta_{X_j}, \beta_{(X_j,Pa(X_j))}, K_{X_j^2|Pa(X_j)}) \\ &= \ln \prod_{i=1}^N \prod_{X_j \in X} P_{X_j|Pa(X_j)}(x_j, Pa(x_j)) \\ &= \ln \prod_{i=1}^N \prod_{X_j \in X} (2\pi)^{-\frac{1}{2}} K_{X_j^2|Pa(X_j)}^{-\frac{1}{2}} \exp\left\{-\frac{(x_{ij} - \beta_{(X_j,Pa(X_j))}) \cdot Pa(x_{ij}) - \beta_{X_j})^2}{2 \cdot K_{X_j^2|Pa(X_j)}}\right\} \\ &= \sum_{X_j \in X} \sum_{i=1}^N -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(K_{X_j^2|Pa(X_j)}) - \frac{1}{2K_{X_j^2|Pa(X_j)}} \cdot (x_{ij} - \beta_{(X_j,Pa(X_j))}) \cdot Pa(x_{ij}) - \beta_{X_j})^2 \\ &= \sum_{X_j \in X} -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \cdot \ln(K_{X_j^2|Pa(X_j)}) - \frac{1}{2K_{X_j^2|Pa(X_j)}} \cdot \sum_{i=1}^N (x_{ij} - \beta_{(X_j,Pa(X_j))}) \cdot Pa(x_{ij}) - \beta_{X_j})^2 \\ &= \sum_{X_j \in X} \frac{-N}{2} \ln(2\pi \cdot K_{X_j^2|Pa(X_j)}) - \frac{1}{2K_{X_j^2|Pa(X_j)}} \cdot \sum_{i=1}^N (x_{ij} - \beta_{(X_j,Pa(X_j))}) \cdot Pa(x_{ij}) - \beta_{X_j})^2 \\ &= \sum_{X_j \in X} \frac{-N}{2} \ln(2\pi \cdot K_{X_j^2|Pa(X_j)}) - \frac{SSR(X_j, Pa(X_j))}{2K_{X_j^2|Pa(X_j)}} \end{aligned}$$

We know that, for a population of size N , we have the following equalities:

$$K_{X_j^2|Pa(X_j)} = K_{X_j^2} - K_{X_j, Pa(X_j)} \cdot K_{Pa^2(X_j)}^{-1} \cdot K_{Pa(X_j)X_j} = \frac{SSR(X_j, Pa(X_j))}{N} = \frac{\sum_{i=1}^N (x_{ij} - \beta_{(X_j, Pa(X_j))} \cdot Pa(x_{ij}) - \beta_{X_j})^2}{N}$$

The likelihood function will therefore only depend on the slopes $\beta_{(X_j, Pa(X_j))}$ and the intercept coefficients β_{X_j} :

$$\bar{L}(D|\beta_{X_j}, \beta_{(X_j, Pa(X_j))}) = \sum_{X_j \in X} \frac{N}{2} \ln(2\pi \cdot K_{X_j^2|Pa(X_j)}) + \sum_{X_j \in X} \frac{SSR(X_j, Pa(X_j))}{2K_{X_j^2|Pa(X_j)}}$$

Using the relations:

$$\begin{aligned} & \sum_{X_j \in X} \frac{N}{2} \ln(2\pi \cdot K_{X_j^2|Pa(X_j)}) \\ &= \sum_{X_j \in X} \frac{N}{2} \ln(2\pi \cdot e \cdot K_{X_j^2|Pa(X_j)}) - \frac{n \cdot N}{2} \\ &= N \cdot \sum_{X_j \in X} h(X_j|Pa(X_j)) - \frac{n \cdot N}{2} \end{aligned}$$

and

$$\sum_{X_j \in X} \frac{SSR(X_j|Pa(X_j))}{2K_{X_j^2|Pa(X_j)}} = \frac{n \cdot N}{2}$$

we obtain:

$$\bar{L}(D|\beta_{X_j}, \beta_{(X_j, Pa(X_j))}) = N \cdot \sum_{X_j \in X} h(X_j|Pa(X_j))$$

From the likelihood function and the number of slopes and intercept coefficients equal to $q + n = \#\{\beta_{X_j, P(X_j)}, \beta_{X_j}\}_{j=1,2,\dots,n}$, we can build the BIC score:

$$BIC(D|\mathcal{B}) = \bar{L}(D|\beta_{X_j}, \beta_{(X_j, Pa(X_j))}) + \frac{(q+n) \ln(N)}{2} = N \cdot \sum_{X_j \in X} h(X_j|Pa(X_j)) + \frac{(q+n) \ln(N)}{2}$$

where:

$\sum_{X_j \in X} h(X_j|Pa(X_j))$ is the Bayesian network's entropy.

q is the number of edges (number of slopes) in the Bayesian network .

n is the number of nodes in the Bayesian network.

N is the number of points (number of rows) of the data frame D .

5.1 Algorithm for Learning data frame from Bayesian Networks

The goal of the algorithm will be to consider a large number of topological sorts linked to chain rules having an entropy equal to $h(X|\mathcal{B}^C)$ (Note that we have $n!$ possible chain rules for a data frame of size $N \times n$). For a fixed topological order, it will be necessary to iteratively remove the edges producing the smallest entropy variations of Bayesian network. After removing the edges, will remain the Bayesian network without edges having an entropy equal to $h(X|\mathcal{B}^R)$.

At each iteration, we will evaluate the Bayesian network's entropy $h(X|\mathcal{B})$ which will vary and increase between the following bounds:

$$h(X|\mathcal{B}^C) \leq h(X|\mathcal{B}) \leq h(X|\mathcal{B}^R)$$

we will compute for each iterated the BIC score:

$$BIC(D|\mathcal{B}) = \bar{L}(D|\beta_{X_j}, \beta_{(X_j, Pa(X_j))}) + \frac{(q+n) \ln(N)}{2} = N \cdot \sum_{X_j \in X} h(X_j|Pa(X_j)) + \frac{(q+n) \ln(N)}{2}$$

The most likely optimal Bayesian network \mathcal{B}^* will be the network that minimizes the following optimization problem:

$$\min_{\mathcal{B}^*} BIC(D|\mathcal{B})$$

In the example considered to apply the learning algorithm, we will set a topological order linked to a chain rule. We will show step by step the iterates leading to the optimal Bayesian network. The example will be a data frame containing car brands.

5.2 Continuous data inference from a Bayesian network

We want to infer a continuous data matrix of size $N \times n$ in topological order from the graph related to the Bayesian network. This is achieved by using OLS (Ordinary least squares):

$$x_{ij} = \sum_{y_{ij} \in Pa(x_{ij})} \beta_{ij} y_{ij} + \beta_j \quad \text{for } i=1,2,\dots,N$$

to which we add gaussian random column vectors with zero mean and a conditional variance $K_{X_j^2|Pa(X_j)}$

6 Learning a data frame from Bayesian network

X1	X2	X3	X4	X5	X6
21.697356	212.496303	100.27983	4.067217	3.1128370	20.45330
17.933487	334.547171	216.25136	3.032607	3.9452276	17.51779
22.593178	293.789279	131.11323	3.847017	2.8745655	17.27431
34.049362	-140.459877	59.76671	5.185856	0.3781136	18.91340
18.893331	193.854070	165.98525	3.619441	2.8882926	17.09582
27.386443	183.449699	89.49365	4.120053	2.2747225	19.61326
29.387658	-27.047273	48.51673	4.005462	1.2374293	19.60590
13.803899	289.576913	203.10891	2.691737	4.4497561	17.63801
23.307997	190.350364	83.14425	4.433126	2.3252469	21.52021
34.096057	2.741246	47.94401	3.978904	0.8459643	16.59992
19.337734	229.179303	148.90931	3.141172	3.2224283	18.05137
12.740558	332.567200	198.15682	3.502937	3.5581512	17.08335
19.019523	177.643152	75.71239	3.984464	3.0206339	19.12922
14.515920	251.345140	238.63220	3.392142	3.8888359	15.05471
24.641912	156.073251	172.47024	3.922760	2.1692936	17.03247
22.308028	5.969799	118.17383	4.371926	1.4265646	18.26750
17.009185	351.668352	214.58385	2.698569	4.2832532	17.28782
19.228647	256.121885	158.85563	4.233123	3.0393819	17.76284
25.065331	192.011334	184.91772	3.628895	2.3155048	18.26881
26.441899	158.193829	115.22245	4.897830	2.2291997	17.05538
24.921998	72.476092	79.17064	3.897563	2.2206522	18.50567
9.768450	360.315682	190.86103	2.513282	4.2600833	19.36172
21.708682	230.343087	148.22627	3.872064	2.9246166	18.02496
23.293301	187.338921	137.30990	3.753809	2.9361905	16.60604
24.776262	102.153614	141.87334	3.991304	2.1954443	16.27751
17.292975	209.795454	103.01824	3.078183	3.4375820	18.70655
20.456419	148.953697	142.97339	3.878322	2.1425225	17.52720
20.389620	186.221825	172.58877	3.685490	2.6114183	17.69306
24.153918	88.997109	73.04434	3.855536	2.7443201	21.19175
12.366297	345.902899	223.52569	2.889015	3.6826036	16.73392
13.257675	499.108686	204.25494	2.968723	4.7720203	19.52203
22.299240	89.665515	122.32289	3.868410	2.5990239	18.06212
8.903630	459.562948	255.89935	3.511985	4.6287205	16.20319
16.803197	339.644711	156.85455	3.236518	3.9164802	16.86312
23.318325	218.850544	121.29319	3.717060	2.9066304	17.16336
19.983920	217.473115	149.56058	3.144869	3.5337949	17.99537
20.879636	175.264826	189.29793	3.727292	2.5445839	17.28957
13.007037	328.143139	188.78226	2.739336	4.3101301	18.32834
19.705524	255.280329	166.26701	3.927320	3.3045594	19.21580
12.909625	371.599198	208.89409	2.813123	4.3670589	20.09771
26.607515	-11.397382	90.50657	4.801016	1.7515093	18.51810
18.357485	273.627256	110.10703	3.721136	3.5930101	19.25385
21.734258	194.486630	147.25357	4.265773	2.7584258	15.87150
18.679990	166.121706	153.32705	3.813516	2.9196218	17.03728
18.013215	281.022416	170.99185	3.164209	3.6305637	19.03299
25.210935	156.325110	108.98786	4.197637	1.9370005	18.86877
25.474886	120.080876	86.99225	3.575362	3.1195229	19.01564
25.558855	149.638937	186.34757	4.247228	2.3641699	14.06583
19.093939	241.095004	123.48452	183.460038	3.6587681	18.50821
26.383655	144.062768	114.88925	4.624856	2.2478064	18.06069
16.631879	415.350156	202.13384	3.246128	4.3677528	18.03676
17.429027	499.253634	227.75786	3.007525	5.0938728	16.08398

X1	X2	X3	X4	X5	X6
16.624258	382.803044	192.96049	2.737288	4.5662243	15.58820
16.066352	306.918324	206.62547	3.236338	3.5222393	15.67147
17.504180	219.025422	209.76230	3.814828	3.4294919	16.88691
21.911348	211.557889	78.55344	4.055491	2.9154360	18.01431
26.577185	144.955804	113.40457	4.097286	2.5707701	17.95062
10.088415	358.589760	231.47066	3.259937	4.1975334	17.73663
23.432039	147.405877	62.37464	3.093330	3.4331099	21.53923
8.573387	424.640718	205.03377	3.087174	4.9942198	17.25097
16.592986	232.340174	119.36525	2.995524	4.2073326	18.54980
28.562558	136.617252	90.23244	4.337025	2.1406164	17.13835
27.511746	165.838291	67.28764	3.941838	1.6319913	16.21451
24.109918	202.000099	193.57244	4.109444	3.1691412	16.63171
15.224393	439.424883	202.70133	3.251049	3.9443693	17.80867
25.981988	200.585319	138.01604	3.728482	3.4324030	19.31789
22.375536	138.939767	120.75469	3.566029	3.3694704	18.77820
18.302730	312.679313	273.84440	3.401653	3.9555278	15.35012
16.496028	278.680945	126.55229	4.185533	3.1401651	18.77259
10.648459	330.025619	237.83779	3.005619	4.2221736	18.54193
19.577220	318.735572	176.17215	3.228952	3.3865012	17.64093
12.602861	547.610755	325.18760	3.364679	4.5170861	16.94013
23.483835	269.181617	221.92444	3.724709	2.8467341	13.64431
11.520018	436.855756	264.40488	3.735643	4.7439170	15.15520
22.735042	239.556194	103.32283	4.569178	2.6527874	17.80767
32.766632	84.141232	54.03472	4.122624	1.3440861	19.77657
3.368601	560.702187	276.03769	2.742008	5.5880158	16.24667
18.582352	223.744131	92.24361	3.101948	3.8218877	19.78210
25.259709	142.345346	102.57292	4.157116	2.5808681	18.65086
25.862437	91.731591	115.25040	3.725904	2.9508191	18.51111
24.405828	264.105590	128.24257	3.379417	3.7308016	17.89932
26.484348	198.879063	195.17488	3.908468	2.1831227	13.83249
27.577924	131.411099	88.60592	3.664482	2.3816511	19.36957
16.500382	320.080511	143.27535	3.873874	3.9361777	18.36889
19.695883	197.885483	122.45748	3.494311	2.8443474	16.61144
18.822765	336.902803	192.74357	3.289000	3.7248344	16.63973
23.940722	206.683049	157.44793	3.657017	3.1324352	19.14949
11.244597	445.089964	244.00301	3.483708	4.5950475	17.16837
12.142727	446.114532	198.29972	3.270059	4.5206776	16.62513
17.717054	262.381335	158.13466	3.228514	4.1851491	19.39476
27.938884	122.305509	23.85977	3.736659	3.0321913	22.76239
23.992895	88.638377	66.26240	4.009436	2.6843940	18.06440
18.545182	309.421665	150.63214	3.405085	3.3540838	19.26326
29.511690	189.962882	88.99168	4.235150	2.0802414	19.23400
24.473435	157.837949	158.95482	4.284887	2.1564811	13.37018
15.336709	307.162666	206.02291	3.002996	3.2221951	18.15282
36.603287	-52.933877	-54.60674	4.250180	1.3984221	23.06948
20.055056	238.642997	164.02440	3.947206	2.8032071	16.87895
21.648469	188.946992	138.74106	3.861113	2.8219705	18.50646
17.156453	353.615611	121.96272	4.664249	3.8069832	19.12094

This data frame has 6 columns and $N = 100$ rows. In this example, we consider the topological sort $(X_5, X_2, X_4, X_3, X_1, X_6)$ and we use the notation \tilde{X}_j to remove the node X_j , the algorithm begin with the following chain rule:

$$\begin{aligned} & h(X_5) + h(X_2|X_5) + h(X_4|X_5X_2) + h(X_3|X_5X_2X_4) + h(X_1|X_5X_2X_4X_3) + h(X_6|X_5X_2X_4X_3X_1) \\ &= 1.399663 + 5.472459 + 0.3482677 + 5.057505 + 2.354467 + 1.587467 \\ &= 16.21983 \end{aligned}$$

$$BIC(D|\mathcal{B}) = N \cdot \sum_{X_j \in \mathcal{X}} h(X_j|Pa(X_j)) + \frac{(q+n)\ln(N)}{2} = 1621.983 + \frac{(15+6)\ln(100)}{2} = 1670.337$$

$$\begin{aligned} & h(X_5) + h(X_2|X_5) + h(X_4|X_5X_2) + h(X_3|X_5X_2X_4) + h(X_1|X_5X_2X_4X_3) + h(X_6|X_5\tilde{X}_2X_4X_3X_1) \\ &= 1.399663 + 5.472459 + 0.3482677 + 5.057505 + 2.354467 + 1.587633 \\ &= 16.21999 \end{aligned}$$

$$BIC(D|\mathcal{B}) = N \cdot \sum_{X_j \in \mathcal{X}} h(X_j|Pa(X_j)) + \frac{(q+n)\ln(N)}{2} = 1621.999 + \frac{(14+6)\ln(100)}{2} = 1668.051$$

$$\begin{aligned} & h(X_5) + h(X_2|X_5) + h(X_4|X_5X_2) + h(X_3|X_5X_2X_4) + h(X_1|X_5X_2\tilde{X}_4X_3) + h(X_6|X_5\tilde{X}_2X_4X_3X_1) \\ &= 1.399663 + 5.472459 + 0.3482677 + 5.0575505 + 2.356048 + 1.587633 \\ &= 16.22162 \end{aligned}$$

$$BIC(D|\mathcal{B}) = N \cdot \sum_{X_j \in \mathcal{X}} h(X_j|Pa(X_j)) + \frac{(q+n)\ln(N)}{2} = 1622.162 + \frac{(13+6)\ln(100)}{2} = 1665.911$$

$$\begin{aligned} & h(X_5) + h(X_2|X_5) + h(X_4|X_5X_2) + h(X_3|\tilde{X}_5X_2X_4) + h(X_1|X_5X_2\tilde{X}_4X_3) + h(X_6|X_5\tilde{X}_2X_4X_3X_1) \\ &= 1.399663 + 5.472459 + 0.3482677 + 5.06061 + 2.356048 + 1.587633 \\ &= 16.22468 \end{aligned}$$

$$BIC(D|\mathcal{B}) = N \cdot \sum_{X_j \in \mathcal{X}} h(X_j|Pa(X_j)) + \frac{(q+n)\ln(N)}{2} = 1622.468 + \frac{(12+6)\ln(100)}{2} = 1663.915$$

$$\begin{aligned} & h(X_5) + h(X_2|X_5) + h(X_4|X_5X_2) + h(X_3|\tilde{X}_5X_2\tilde{X}_4) + h(X_1|X_5X_2\tilde{X}_4X_3) + h(X_6|X_5\tilde{X}_2X_4X_3X_1) \\ &= 1.399663 + 5.472459 + 0.3482677 + 5.062763 + 2.356048 + 1.587633 \\ &= 16.22683 \end{aligned}$$

$$BIC(D|\mathcal{B}) = N \cdot \sum_{X_j \in \mathcal{X}} h(X_j|Pa(X_j)) + \frac{(q+n)\ln(N)}{2} = 1622.683 + \frac{(11+6)\ln(100)}{2} = 1661.827$$

$$\begin{aligned}
& h(X_5) + h(X_2|X_5) + h(X_4|X_5X_2) + h(X_3|\tilde{X}_5X_2\tilde{X}_4) + h(X_1|X_5X_2\tilde{X}_4X_3) + h(X_6|X_5\tilde{X}_2\tilde{X}_4X_3X_1) \\
& = 1.399663 + 5.472459 + 0.3482677 + 5.062763 + 2.356048 + 1.593285 \\
& = 16.23249
\end{aligned}$$

$$BIC(D|\mathcal{B}) = N \cdot \sum_{X_j \in \mathcal{X}} h(X_j|Pa(X_j)) + \frac{(q+n)\ln(N)}{2} = 1623.249 + \frac{(10+6)\ln(100)}{2} = 1660.09$$

$$\begin{aligned}
& h(X_5) + h(X_2|X_5) + h(X_4|X_5X_2) + h(X_3|\tilde{X}_5X_2\tilde{X}_4) + h(X_1|X_5\tilde{X}_2\tilde{X}_4X_3) + h(X_6|X_5\tilde{X}_2\tilde{X}_4X_3X_1) \\
& = 1.399663 + 5.472459 + 0.3482677 + 5.062763 + 2.361169 + 1.593285 \\
& = 16.23761
\end{aligned}$$

$$BIC(D|\mathcal{B}) = N \cdot \sum_{X_j \in \mathcal{X}} h(X_j|Pa(X_j)) + \frac{(q+n)\ln(N)}{2} = 1623.761 + \frac{(9+6)\ln(100)}{2} = 1658.299$$

$$\begin{aligned}
& h(X_5) + h(X_2|X_5) + h(X_4|X_5\tilde{X}_2) + h(X_3|\tilde{X}_5X_2\tilde{X}_4) + h(X_1|X_5\tilde{X}_2\tilde{X}_4X_3) + h(X_6|X_5\tilde{X}_2\tilde{X}_4X_3X_1) \\
& = 1.399663 + 5.472459 + 0.3565913 + 5.062763 + 2.361169 + 1.593285 \\
& = 16.24593
\end{aligned}$$

$$BIC(D|\mathcal{B}) = N \cdot \sum_{X_j \in \mathcal{X}} h(X_j|Pa(X_j)) + \frac{(q+n)\ln(N)}{2} = 1624.593 + \frac{(8+6)\ln(100)}{2} = 1656.829$$

$$\begin{aligned}
& h(X_5) + h(X_2|X_5) + h(X_4|X_5\tilde{X}_2) + h(X_3|\tilde{X}_5X_2\tilde{X}_4) + h(X_1|X_5\tilde{X}_2\tilde{X}_4X_3) + h(X_6|X_5\tilde{X}_2\tilde{X}_4X_3\tilde{X}_1) \\
& = 1.399663 + 5.472459 + 0.3565913 + 5.062763 + 2.361169 + 1.600407 \\
& = 16.25305
\end{aligned}$$

$$BIC(D|\mathcal{B}) = N \cdot \sum_{X_j \in \mathcal{X}} h(X_j|Pa(X_j)) + \frac{(q+n)\ln(N)}{2} = 1625.305 + \frac{(7+6)\ln(100)}{2} = 1655.239$$

$$\begin{aligned}
& h(X_5) + h(X_2|X_5) + h(X_4|X_5\tilde{X}_2) + h(X_3|\tilde{X}_5X_2\tilde{X}_4) + h(X_1|X_5\tilde{X}_2\tilde{X}_4\tilde{X}_3) + h(X_6|X_5\tilde{X}_2\tilde{X}_4X_3\tilde{X}_1) \\
& = 1.399663 + 5.472459 + 0.3565913 + 5.062763 + 2.483216 + 1.600407 \\
& = 16.3751
\end{aligned}$$

$$BIC(D|\mathcal{B}) = N \cdot \sum_{X_j \in \mathcal{X}} h(X_j|Pa(X_j)) + \frac{(q+n)\ln(N)}{2} = 1637.51 + \frac{(6+6)\ln(100)}{2} = 1665.141$$

$$h(X_5) + h(X_2|X_5) + h(X_4|X_5\tilde{X}_2) + h(X_3|\tilde{X}_5X_2\tilde{X}_4) + h(X_1|X_5\tilde{X}_2\tilde{X}_4\tilde{X}_3) + h(X_6|\tilde{X}_5\tilde{X}_2\tilde{X}_4X_3\tilde{X}_1)$$

$$\begin{aligned}
&= 1.399663 + 5.472459 + 0.3565913 + 5.062763 + 2.483216 + 1.732874 \\
&= 16.50757
\end{aligned}$$

$$BIC(D|\mathcal{B}) = N \cdot \sum_{X_j \in \mathcal{X}} h(X_j|Pa(X_j)) + \frac{(q+n)\ln(N)}{2} = 1650.757 + \frac{(5+6)\ln(100)}{2} = 1676.085$$

$$\begin{aligned}
&h(X_5) + h(X_2|X_5) + h(X_4|X_5\tilde{X}_2) + h(X_3|\tilde{X}_5X_2\tilde{X}_4) + h(X_1|X_5\tilde{X}_2\tilde{X}_4\tilde{X}_3) + h(X_6|\tilde{X}_5\tilde{X}_2\tilde{X}_4\tilde{X}_3\tilde{X}_1) \\
&= 1.399663 + 5.472459 + 0.3565913 + 5.062763 + 2.483216 + 1.949459 \\
&= 16.72415
\end{aligned}$$

$$BIC(D|\mathcal{B}) = N \cdot \sum_{X_j \in \mathcal{X}} h(X_j|Pa(X_j)) + \frac{(q+n)\ln(N)}{2} = 1672.415 + \frac{(4+6)\ln(100)}{2} = 1695.441$$

$$\begin{aligned}
&h(X_5) + h(X_2|X_5) + h(X_4|\tilde{X}_5\tilde{X}_2) + h(X_3|\tilde{X}_5X_2\tilde{X}_4) + h(X_1|X_5\tilde{X}_2\tilde{X}_4\tilde{X}_3) + h(X_6|\tilde{X}_5\tilde{X}_2\tilde{X}_4\tilde{X}_3\tilde{X}_1) \\
&= 1.399663 + 5.472459 + 0.7943387 + 5.062763 + 2.483216 + 1.949459 \\
&= 17.1619
\end{aligned}$$

$$BIC(D|\mathcal{B}) = N \cdot \sum_{X_j \in \mathcal{X}} h(X_j|Pa(X_j)) + \frac{(q+n)\ln(N)}{2} = 1716.19 + \frac{(3+6)\ln(100)}{2} = 1736.913$$

$$\begin{aligned}
&h(X_5) + h(X_2|X_5) + h(X_4|\tilde{X}_5\tilde{X}_2) + h(X_3|\tilde{X}_5\tilde{X}_2\tilde{X}_4) + h(X_1|X_5\tilde{X}_2\tilde{X}_4\tilde{X}_3) + h(X_6|\tilde{X}_5\tilde{X}_2\tilde{X}_4\tilde{X}_3\tilde{X}_1) \\
&= 1.399663 + 5.472459 + 0.7943387 + 5.550476 + 2.483216 + 1.949459 \\
&= 17.64961
\end{aligned}$$

$$BIC(D|\mathcal{B}) = N \cdot \sum_{X_j \in \mathcal{X}} h(X_j|Pa(X_j)) + \frac{(q+n)\ln(N)}{2} = 1764.961 + \frac{(2+6)\ln(100)}{2} = 1783.382$$

$$\begin{aligned}
&h(X_5) + h(X_2|X_5) + h(X_4|\tilde{X}_5\tilde{X}_2) + h(X_3|\tilde{X}_5X_2\tilde{X}_4) + h(X_1|\tilde{X}_5\tilde{X}_2\tilde{X}_4\tilde{X}_3) + h(X_6|\tilde{X}_5\tilde{X}_2\tilde{X}_4\tilde{X}_3\tilde{X}_1) \\
&= 1.399663 + 5.472459 + 0.7943387 + 5.550476 + 3.210796 + 1.949459 \\
&= 18.37719
\end{aligned}$$

$$BIC(D|\mathcal{B}) = N \cdot \sum_{X_j \in \mathcal{X}} h(X_j|Pa(X_j)) + \frac{(q+n)\ln(N)}{2} = 1837.719 + \frac{(1+6)\ln(100)}{2} = 1853.837$$

$$\begin{aligned}
& h(X_5) + h(X_2|\tilde{X}_5) + h(X_4|\tilde{X}_5\tilde{X}_2) + h(X_3|\tilde{X}_5\tilde{X}_2\tilde{X}_4) + h(X_1|\tilde{X}_5\tilde{X}_2\tilde{X}_4\tilde{X}_3) + h(X_6|\tilde{X}_5\tilde{X}_2\tilde{X}_4\tilde{X}_3\tilde{X}_1) \\
& = 1.399663 + 6.270549 + 0.7943387 + 5.550476 + 3.210796 + 1.949459 \\
& = 1917.528
\end{aligned}$$

$$BIC(D|\mathcal{B}) = N \cdot \sum_{X_j \in \mathcal{X}} h(X_j|Pa(X_j)) + \frac{(q+n)\ln(N)}{2} = 1917.528 + \frac{(0+6)\ln(100)}{2} = 1931.344$$

We will represent the graph of the BIC score according to the number of edges (number of slopes)

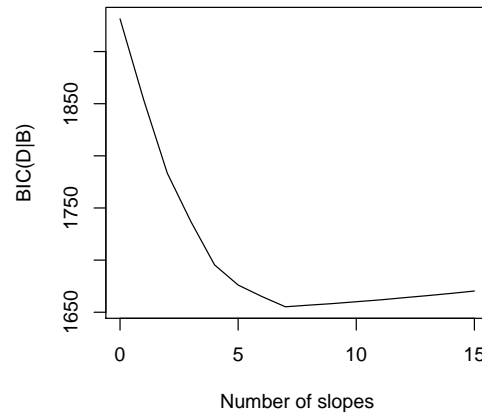


Figure 1: BIC score and number of slopes

We see that the BIC score has a minimum value:

$$\min_{\mathcal{B}^*} BIC(D|\mathcal{B}) = 1655.239$$

\mathcal{B}^* is the most likely optimal Bayesian network for the data frame D.

We will now represent the most likely optimal Bayesian network \mathcal{B}^* :

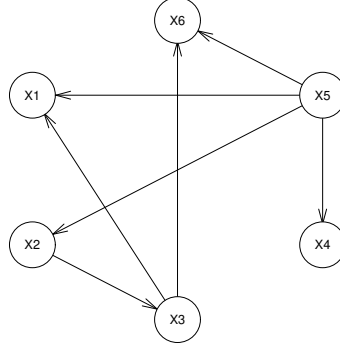


Figure 2: Bayesian network

$$\begin{aligned}
 & h(X_5) + h(X_2|X_5) + h(X_4|X_5\tilde{X}_2) + h(X_3|\tilde{X}_5X_2\tilde{X}_4) + h(X_1|X_5\tilde{X}_2\tilde{X}_4X_3) + h(X_6|X_5\tilde{X}_2\tilde{X}_4X_3\tilde{X}_1) \\
 &= 1.399663 + 5.472459 + 0.3565913 + 5.062763 + 2.361169 + 1.600407 \\
 &= 16.25305
 \end{aligned}$$

$$BIC(D|\mathcal{B}) = N \cdot \sum_{X_j \in \mathcal{X}} h(X_j|Pa(X_j)) + \frac{(q+n)\ln(N)}{2} = 1625.305 + \frac{(7+6)\ln(100)}{2} = 1655.239$$

7 Conclusion

In this paper, we have proposed an algorithm for learning a data frame from Bayesian networks. From a data frame example, we have shown the steps leading to obtaining the most likely optimal Bayesian network.

8 Appendix

Theorem *If K is a symmetric semidefinite matrix of size $n \times n$ then the determinant of this matrix verifies the following inequalities :*

$$0 \leq \det(K) \leq \prod_{i=1}^n K_{ii}$$

where K_{ii} are the diagonal elements of the matrix K

Corollary *If K is a correlation matrix of size $n \times n$ then the determinant of this matrix verifies the following inequalities :*

$$0 \leq \det(K) \leq 1$$

[1] *Elements of information theory*. Author: Thomas M.Cover and Joy A.Thomas. Copyright 1991 John Wiley and sons.

[2] *Optimal stastical decisions*. Author: Morris H.DeGroot. Copyright 1970-2004 John Wiley and sons.

[3] *Matrix Analysis*. Author: Roger A.Horn and Charles R.Johnson. Copyright 2012, Cambridge university press.

[4] *Causality: Models, reasoning and inference*. Author: Judea Pearl .Copyright 2000, Cambridge university press.