Information theory applied to Bayesian network for learning continuous data matrix

Ait-Taleb Nabil*

August 21, 2021

Abstract

In this article, we are proposing a learning algorithm for continuous data matrix based on entropy absorption of a Bayesian network.This method consists in losing a little bit of likelihood compared to a chain rule's best likelihood, in order to get a good idea of the higher conditionings that are taking place between the Bayesian network's nodes.We are presenting the known results related to information theory, the multidimensional Gaussian probability, AIC and BIC scores for continuous data matrix learning from a Bayesian network, and we are showing the entropy absorption algorithm using the Kullback-leibler divergence with an example of continuous data matrix.

*Corresponding author: nabiltravail1982@gmail.com

1 Introduction

In this paper, we will cover information theory for continuous data like differential entropy, joint differential entropy, conditional differential entropy, mutual information, conditional mutual information and the Kullback-leibler divergence. We will make a brief reminder on the Gaussian multidimensional probability and the information theory. We will demonstrate a theorem on conditional entropy inequalities for Gaussian random vectors, this theorem will be later used to bound Bayesian network's differential entropy. In the following, we will define a Bayesian network using a Gaussian random vector, we will show how to compute a Bayesian network's differential entropy and conclude by proposing a theorem to upper and lower bound this differential entropy. In order to do data learning, we will detail, for a Bayesian network the AIC and the BIC scores and a method of differential entropy absorption of a Bayesian network. The differential entropy absorption method will use Kullback-leibler divergence to show the increase in entropy when choosing a Bayesian network model. We will also show how to infer data from a Bayesian network. From an example, this paper will conclude by suggesting a learning algorithm for continuous data matrix based on the differential entropy absorption of a Bayesian network.

2 Information and 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:*

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

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 sizes n and m respectively, with a multidimensional probability density function (pdf) $p_{X_1X_2}(\vec{x}_1, \vec{x}_2)$, we define the joint differential entropy $h(X_1, X_2)$ as:*

$$h(X_1, X_2) = -\int_{\mathbb{X}_1} \int_{\mathbb{X}_2} p_{X_1X_2}(\vec{x}_1, \vec{x}_2) \ln p_{X_1X_2}(\vec{x}_1, \vec{x}_2) \overrightarrow{dx_1} \overrightarrow{dx_2}$$

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 sizes n et m respectively, with a multidimensional probability density function (pdf) $p_{X_1X_2}(\vec{x}_1, \vec{x}_2)$, we define the conditional differential entropy $h(X_1|X_2)$ as:*

$$h(X_1|X_2) = -\int_{\mathbb{X}_1} \int_{\mathbb{X}_2} p_{X_1X_2}(\vec{x}_1, \vec{x}_2) \ln p_{X_1|X_2}(\vec{x}_1, \vec{x}_2) \overrightarrow{dx_1} \overrightarrow{dx_2}$$

where we have:

$$p_{X_1|X_2}(\vec{x}_1, \vec{x}_2) = \frac{p_{X_1X_2}(\vec{x}_1, \vec{x}_2)}{p_{X_2}(\vec{x}_2)}$$

$$P_{X_2}(\vec{x}_2) = \int_{\mathbb{X}_1} p_{X_1X_2}(\vec{x}_1, \vec{x}_2) \overrightarrow{dx_1}$$

2.4 Joint differential entropy and conditional differential entropy

Given two concatenated random vectors $(\vec{x}_1, \vec{x}_2)$, defined on the sets $\mathbb{X}_1$ and $\mathbb{X}_2$ of sizes n et m respectively, with a multidimensional probability density function (pdf) $p_{X_1X_2}(\vec{x}_1, \vec{x}_2)$, we can then establish the relation between joint differential entropy and conditional differential entropy as:

$$h(X_1|X_2) = h(X_1, X_2) - h(X_2)$$

Indeed:

$$\begin{aligned}
&h(X_1|X_2)\\
&= -\int_{\mathbb{X}_1}\int_{\mathbb{X}_2} p_{X_1X_2}(\vec{x}_1, \vec{x}_2)\ln p_{X_1|X_2}(\vec{x}_1|\vec{x}_2)\overrightarrow{dx_1}\overrightarrow{dx_2}\\
&= -\int_{\mathbb{X}_1}\int_{\mathbb{X}_2} p_{X_1X_2}(\vec{x}_1, \vec{x}_2)\ln \frac{p_{X_1X_2}(\vec{x}_1, \vec{x}_2)}{p_{X_2}(\vec{x}_2)}\overrightarrow{dx_1}\overrightarrow{dx_2}\\
&= -\int_{\mathbb{X}_1}\int_{\mathbb{X}_2} P_{X_1X_2}(\vec{x}_1, \vec{x}_2)\ln P_{X_1X_2}(\vec{x}_1, \vec{x}_2)\overrightarrow{dx_1}\overrightarrow{dx_2} + \int_{\mathbb{X}_2}(\int_{\mathbb{X}_1} p_{X_1X_2}(\vec{x}_1, \vec{x}_2)\overrightarrow{dx_1})\ln P_{X_2}(\vec{x}_2)\overrightarrow{dx_2}\\
&= -\int_{\mathbb{X}_1}\int_{\mathbb{X}_2} P_{X_1X_2}(\vec{x}_1, \vec{x}_2)\ln P_{X_1X_2}(\vec{x}_1, \vec{x}_2)\overrightarrow{dx_1}\overrightarrow{dx_2} + \int_{\mathbb{X}_2} P_{X_2}(\vec{x}_2)\ln P_{X_2}(\vec{x}_2)\overrightarrow{dx_2}\\
&= h(X_1, X_2) - h(X_2)
\end{aligned}$$

2.5 Mutual information between 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 sizes n et m respectively, with a multidimensional probability density function (pdf) $p_{X_1X_2}(\vec{x}_1, \vec{x}_2)$, we define the mutual information $I(X_1, X_2)$ between two random vectors as:*

$$I(X_1, X_2) = \int_{\mathbb{X}_1}\int_{\mathbb{X}_2} p_{X_1X_2}(\vec{x}_1, \vec{x}_2)\ln \frac{p_{X_1X_2}(\vec{x}_1, \vec{x}_2)}{P_{X_1}(x_1)p_{X_2}(x_2)}\overrightarrow{dx_1}\overrightarrow{dx_2}$$

We can make the link between mutual information and the differential entropy:

$$\begin{aligned}
I(X_1, X_2) &= \int_{\mathbb{X}_1}\int_{\mathbb{X}_2} p_{X_1X_2}(\vec{x}_1, \vec{x}_2)\ln \frac{p_{X_1X_2}(\vec{x}_1, \vec{x}_2)}{P_{X_1}(x_1)p_{X_2}(x_2)}\overrightarrow{dx_1}\overrightarrow{dx_2}\\
&= \int_{\mathbb{X}_1}\int_{\mathbb{X}_2} p_{X_1X_2}(\vec{x}_1, \vec{x}_2)\ln \frac{P_{X_1|X_2}(\vec{x}_1, \vec{x}_2)}{P_{X_1}(\vec{x}_1)}\overrightarrow{dx_1}\overrightarrow{dx_2}\\
&= -\int_{\mathbb{X}_1}(\int_{\mathbb{X}_2} p_{X_1X_2}(\vec{x}_1, \vec{x}_2)\ln p_{X_1}(\vec{x}_1)\overrightarrow{dx_2})\overrightarrow{dx_1} + \int_{\mathbb{X}_1}\int_{\mathbb{X}_2} p_{X_1X_2}(\vec{x}_1, \vec{x}_2)\ln P_{X_1|X_2}(\vec{x}_1, \vec{x}_2)\overrightarrow{dx_1}\overrightarrow{dx_2}\\
&= -\int_{\mathbb{X}_1} p_{X_1}(\vec{x}_1)\ln p_{X_1}(\vec{x}_1)\overrightarrow{dx_1} + \int_{\mathbb{X}_1}\int_{\mathbb{X}_2} p_{X_1X_2}(\vec{x}_1, \vec{x}_2)\ln P_{X_1|X_2}(\vec{x}_1, \vec{x}_2)\overrightarrow{dx_1}\overrightarrow{dx_2}\\
&= h(X_1) - h(X_1|X_2)
\end{aligned}$$

2.6 Conditional mutual information between two random vectors given a random vector

Definition: *Given three concatenated random vectors* $(\vec{x}_1, \vec{x}_2, \vec{x}_3)$*, defined on the sets* $\mathbb{X}_1$*,* $\mathbb{X}_2$ *and* $\mathbb{X}_3$ *of sizes n ,m and l respectively, with a multidimensional probability density function (pdf)* $p_{X_1X_2X_3}(\vec{x}_1, \vec{x_2}, \vec{x}_3)$*, we define the conditional mutual information* $I(X_1, X_2|X_3)$ *between two random vectors given a random vector as:*

$$I(X_1, X_2|X_3) = \int_{\mathbb{X}_1}\int_{\mathbb{X}_2}\int_{\mathbb{X}_3} p_{X_1X_2X_3}(\vec{x}_1, \vec{x}_2, \vec{x}_3) \ln \frac{p_{X_1,X_2|X_3}(\vec{x}_1, \vec{x}_2, \vec{x}_3)}{p_{X_1|X_3}(\vec{x}_1, \vec{x}_3) p_{X_2|X_3}(\vec{x}_2, \vec{x}_3)} \overrightarrow{dx_1}\overrightarrow{dx_2}\overrightarrow{dx_3}$$

wihch can also be written :

$$I(X_1, X_2|X_3) = \int_{\mathbb{X}_1}\int_{\mathbb{X}_2}\int_{\mathbb{X}_3} p_{X_1,X_2,X_3}(\vec{x}_1, \vec{x}_2, \vec{x}_3) \ln \frac{P_{X_1,X_2,X_3}(\vec{x}_1, \vec{x}_2, \vec{x}_3) p_{X_3}(\vec{x}_3)}{p_{X_1,X_3}(\vec{x}_1, \vec{x}_2) p_{X_2,X_3}(\vec{x_2}, \vec{x}_3)} \overrightarrow{dx_1}\overrightarrow{dx_2}\overrightarrow{dx_3}$$

2.7 Conditional mutual information, the joint and conditional differential entropies

In this section, we will express the conditional mutual information as a function of the joint and conditional differential entropies

$$I(X_1, X_2|X_3)$$

$$= \int_{\mathbb{X}_1}\int_{\mathbb{X}_2}\int_{\mathbb{X}_3} p_{X_1,X_2,X_3}(\vec{x}_1, \vec{x}_2, \vec{x}_3) \ln \frac{P_{X_1,X_2,X_3}(\vec{x}_1,\vec{x}_2,\vec{x}_3) p_{X_3}(\vec{x}_3)}{p_{X_1,X_3}(\vec{x}_1,\vec{x}_2) p_{X_2,X_3}(\vec{x_2},\vec{x}_3)} \overrightarrow{dx_1}\overrightarrow{dx_2}\overrightarrow{dx_3}$$

$$= \int_{\mathbb{X}_1}\int_{\mathbb{X}_2}\int_{\mathbb{X}_3} p_{X_1,X_2,X_3}(\vec{x}_1, \vec{x_2}, \vec{x}_3) \ln p_{X_1,X_2,X_3}(\vec{x}_1, \vec{x_2}, \vec{x}_3) \overrightarrow{dx_1}\overrightarrow{dx_2}\overrightarrow{dx_3}$$

$$+ \int_{\mathbb{X}_3} \{ \int_{\mathbb{X}_1}\int_{\mathbb{X}_2} p_{X_1,X_2,X_3}(\vec{x}_1, \vec{x_2}, \vec{x}_3) \overrightarrow{dx_1}\overrightarrow{dx_2} \} \ln p_{X_3}(\vec{x}_3) \overrightarrow{dx_3}$$

$$- \int_{\mathbb{X}_1}\int_{\mathbb{X}_3} \{ \int_{\mathbb{X}_2} p_{X_1,X_2,X_3}(\vec{x}_1, \vec{x}_2, \vec{x}_3) \overrightarrow{dx_2} \} \ln p_{X_1,X_3}(\vec{x}_1, \vec{x}_3) \overrightarrow{dx_1}\overrightarrow{dx_3}$$

$$- \int_{\mathbb{X}_2}\int_{\mathbb{X}_3} \{ \int_{\mathbb{X}_1} p_{X_1,X_2,X_3}(\vec{x}_1, \vec{x}_2, \vec{x}_3) \overrightarrow{dx_1} \} \ln p_{X_2,X_3}(\vec{x}_2, \vec{x}_3) \overrightarrow{dx_2}\overrightarrow{dx_3}$$

$$= \int_{\mathbb{X}_1}\int_{\mathbb{X}_2}\int_{\mathbb{X}_3} p_{X_1,X_2,X_3}(\vec{x}_1, \vec{x_2}, \vec{x}_3) \ln p_{X_1,X_2,X_3}(\vec{x}_1, \vec{x_2}, \vec{x}_3) \overrightarrow{dx_1}\overrightarrow{dx_2}\overrightarrow{dx_3}$$

$$+ \int_{\mathbb{X}_3} p_{X_3}(\vec{x}_3) \ln p_{X_3}(\vec{x}_3) \overrightarrow{dx_3}$$

$$- \int_{\mathbb{X}_1}\int_{\mathbb{X}_3} p_{X_1,X_3}(\vec{x}_1, \vec{x}_3) \ln p_{X_1,X_3}(\vec{x}_1, \vec{x}_3) \overrightarrow{dx_1}\overrightarrow{dx_3}$$

$$- \int_{\mathbb{X}_2}\int_{\mathbb{X}_3} p_{X_2,X_3}(\vec{x}_2, \vec{x}_3) \ln p_{X_2,X_3}(\vec{x}_2, \vec{x}_3) \overrightarrow{dx_2}\overrightarrow{dx_3}$$

$$= -h(X_1, X_2, X_3) - h(X_3) + h(X_1, X_3) + h(X_2, X_3)$$

$$= h(X_1, X_3) - h(X_3) - h(X_1, X_2, X_3) + h(X_2, X_3)$$

$$= h(X_1|X_3) - h(X_1|X_2X_3)$$

$$= h(X_2|X_3) - h(X_2|X_1X_3)$$

Definition:

The Kullback-Leibler divergence between probability density functions $p(\vec{x})$ *and* $q(\vec{x})$ *defined on the set X is:*

$$D_{KL}(p(\vec{x})||q(\vec{x})) = \int_{\mathbb{X}} p(\vec{x}) \ln(\frac{p(\vec{x})}{q(\vec{x})}) \vec{d}x$$

The Kullback-leibler divergence will be used later for the entropy absorption algorithm.

3 Multivariate Gaussian distribution and information theory

3.1 Joint and conditional gaussian multidimensional probability

Consider a partioned 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 sizes 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}:

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

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:

$W_{X_2^2} = K_{X_2^2}^{-1} + W_{X_2X_1}.W_{X_1^2}^{-1}.W_{X_1X_2}$.

Indeed:

$$K_{X^2} W_{X^2} = \begin{pmatrix} K_{X_1^2} W_{X_1^2} + K_{X_1X_2} W_{X_2X_1} & K_{X_1^2} W_{X_1X_2} + K_{X_1X_2} W_{X_2^2} \\ K_{X_2X_1} W_{X_1^2} + K_{X_2^2} W_{X_2X_1} & K_{X_2X_1} W_{X_1X_2} + K_{X_2^2} W_{X_2^2} \end{pmatrix} = \begin{pmatrix} I_{k_1,k_1} & 0 \\ 0 & I_{k_2,k_2} \end{pmatrix}$$

$$K_{X_2X_1} W_{X_1^2} + K_{X_2^2} W_{X_2X_1} = 0$$

$$K_{X_2^2}^{-1} K_{X_2X_1} W_{X_1^2} + W_{X_2X_1} = 0$$

$$K_{X_2^2}^{-1} K_{X_2X_1} = -W_{X_2X_1} W_{X_1^2}^{-1}$$

$$K_{X_2^2} W_{X_2^2} + K_{X_2X_1} W_{X_1X_2} = I_{k_2k_2}$$

$$W_{X_2^2} = K_{X_2^2}^{-1} - K_{X_2}^{-1}.K_{X_2X_1}.W_{X_1X_2}$$

Finaly, we obtain:

$$W_{X_2^2} = K_{X_2^2}^{-1} + W_{X_2X_1}.W_{X_1^2}^{-1}.W_{X_1X_2}$$

Now, we will develop the Mahalanobis distance:

$(\vec{x}-\vec{\mu_X})^t W_{X^2}(\vec{x}-\vec{\mu_X})$

$= (\vec{x_1}-\vec{\mu_{X_1}}, \vec{x_2}-\vec{\mu_{X_2}}) \begin{pmatrix} W_{X_1^2} & W_{X_1X_2} \\ W_{X_2X_1} & W_{X_2^2} \end{pmatrix} \begin{pmatrix} \vec{x_1}-\vec{\mu_{X_1}} \\ \vec{x_2}-\vec{\mu_{X_2}} \end{pmatrix}$

$= (\vec{x_1}-\vec{\mu_{X_1}})^t W_{X_1^2}(\vec{x_1}-\vec{\mu_{X_1}}) + (\vec{x_1}-\vec{\mu_{X_1}})^t W_{X_1X_2}(\vec{x_2}-\vec{\mu_{X_2}}) + (\vec{x_2}-\vec{\mu_{X_2}})^t W_{X_2X_1}(\vec{x_1}-\vec{\mu_{X_1}})$

$+(\vec{x_2}-\vec{\mu_{X_2}})^t W_{X_2^2}(\vec{x_2}-\vec{\mu_{X_2}})$

Using the relation: $W_{X_2^2} = K_{X_2^2}^{-1} + W_{X_2X_1}.W_{X_1^2}^{-1}.W_{X_1X_2}$, we obtain:

$= (\vec{x_1}-\vec{\mu_{X_1}})^t W_{X_1^2}(\vec{x_1}-\vec{\mu_{X_1}}) + (\vec{x_1}-\vec{\mu_{X_1}})^t W_{X_1X_2}(\vec{x_2}-\vec{\mu_{X_2}}) + (\vec{x_2}-\vec{\mu_{X_2}})^t W_{X_2X_1}(\vec{x_1}-\vec{\mu_{X_1}})$

$+(\vec{x_2}-\vec{\mu_{X_2}})^t W_{X_2X_1} W_{X_1^2}^{-1} W_{X_1X_2}(\vec{x_2}-\vec{\mu_{X_2}}) + (\vec{x_2}-\vec{\mu_{X_2}})^t K_{X_2^2}^{-1}(\vec{x_2}-\vec{\mu_{X_2}})$

$= [(\vec{x_1}-\vec{\mu_{X_1}}) + W_{X_1^2}^{-1} W_{X_1X_2}(\vec{x_2}-\vec{\mu_{X_2}})]^t [W_{X_1^2}(\vec{x_1}-\vec{\mu_{X_1}}) + W_{X_1X_2}(\vec{x_2}-\vec{\mu_{X_2}})]$

$+(\vec{x_2}-\vec{\mu_{X_2}})^t K_{X_2^2}^{-1}(\vec{x_2}-\vec{\mu_{X_2}})$

$= [(\vec{x_1}-\vec{\mu_{X_1}}) + W_{X_1^2}^{-1} W_{X_1X_2}(\vec{x_2}-\vec{\mu_{X_2}})]^t W_{X_1^2}.[(\vec{x_1}-\vec{\mu_{X_1}}) + W_{X_1^2}^{-1} W_{X_1X_2}(\vec{x_2}-\vec{\mu_{X_2}})]$

$+(\vec{x_2}-\vec{\mu_{X_2}})^t K_{X_2^2}^{-1}(\vec{x_2}-\vec{\mu_{X_2}})$

We put:

$Q_1 = (\vec{x_1}-\vec{\nu_{X_1/X_2}})^t (K_{X_1^2} - K_{X_1X_2}K_{X_2^2}^{-1}K_{X_2X_1})^{-1}(\vec{x_1}-\vec{\nu_{X_1/X_2}})$

$\vec{\nu_{X_1|X_2}} = \vec{\mu_{X_1}} + K_{X_1X_2}K_{X_2^2}^{-1}(\vec{x_2}-\vec{\mu_{X_2}})$

$Q_2 = (\vec{x_2}-\vec{\mu_{X_2}})^t K_{X_2^2}^{-1}(\vec{x_2}-\vec{\mu_{X_2}})$

We then obtain the equalities as follows:

$(\vec{x}-\vec{\mu_X})^t K_{X^2}^{-1}(\vec{x}-\mu_X) = Q_1 + Q_2$

$P_X(\vec{x}) = (2\pi)^{-\frac{n}{2}}|K_{X^2}|^{-\frac{1}{2}} \exp\{-\frac{Q_1+Q_2}{2}\} = (2\pi)^{-\frac{n}{2}}|K_{X^2}|^{-\frac{1}{2}} \exp\{(\vec{x}-\vec{\mu_X})^t K_{X^2}^{-1}(\vec{x}-\vec{\mu_X})\}$

$P_{X_2}(\vec{x_2}) = (2\pi)^{-\frac{k_2}{2}}|K_{X_2^2}|^{-\frac{1}{2}} \exp\{-\frac{Q_2}{2}\} = (2\pi)^{-\frac{k_2}{2}}|K_{X_2^2}| \exp\{-\frac{(\vec{x_2}-\vec{\mu_{X_2}})^t K_{X_2^2}^{-1}(\vec{x_2}-\vec{\mu_{X_2}})}{2}\}$

Using the relation $\frac{P_{X_1X_2}(\vec{x_1},\vec{x_2})}{P_{X_2}(x_2)}$:

$P_{X_1|X_2}(x_1,x_2) = (2\pi)^{-\frac{k_1}{2}}(\frac{|K_{(X_1X_2)^2}|}{|K_{X_2^2}|})^{-\frac{1}{2}} \exp\{-\frac{Q_1}{2}\} = (2\pi)^{-\frac{k_1}{2}}(\frac{|K_{(X_1X_2)^2}|}{|K_{X_2^2}|})^{-\frac{1}{2}} \exp\{-\frac{(\vec{x_1}-\vec{\nu_{X_1/X_2}})^t (K_{X_1^2}-K_{X_1X_2}K_{X_2^2}^{-1}K_{X_2X_1})^{-1}(\vec{x_1}-\vec{\nu_{X_1/X_2}})}{2}\}$

If we use the Schur's complement $K_{X_1^2|X_2} = K_{X_1^2} - K_{X_1X_2}K_{X_2^2}^{-1}K_{X_2X_1}$

we can express the conditional probability $P_{X_1|X_2}(\vec{x_1},\vec{x_2})$ as follows:

$P_{X_1|X_2}(\vec{x_1},\vec{x_2})$

$= (2\pi)^{-\frac{k_1}{2}}(\frac{|K_{(X_1X_2)^2}|}{|K_{X_2^2}|})^{-\frac{1}{2}} \exp\{-\frac{(\vec{x_1}-\vec{\nu_{X_1/X_2}})^t K_{X_1^2|X_2}^{-1}(\vec{x_1}-\vec{\nu_{X_1/X_2}})}{2}\} = (2\pi)^{-\frac{k_1}{2}}|K_{X_1^2|X_2}|^{-\frac{1}{2}} \exp\{-\frac{(\vec{x_1}-\vec{\nu_{X_1/X_2}})^t K_{X_1^2|X_2}^{-1}(\vec{x_1}-\vec{\nu_{X_1/X_2}})}{2}\}$

3.2 Differential entropy of a Gaussian random vector

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

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

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:

$h(X)$

$= -\int_{-\infty}^{+\infty} p_X(\vec{x})\ln\{p_X(\vec{x})\}\overrightarrow{dx}$

$= -\int_{-\infty}^{+\infty} p_X(\vec{x})[-\frac{1}{2}(\vec{x} - \mu_X)^t K_{X^2}^{-1}(\vec{x} - \mu_X) - \ln(\sqrt{2\pi})^n|K_{X^2}|^{\frac{1}{2}}]\overrightarrow{dx}$

$= \frac{1}{2}E_X[\sum_{ij}(\vec{x}_i - \mu_{X_i})^t(K_{X^2}^{-1})_{ij}(\vec{x}_j - \mu_{X_j})] + \frac{1}{2}\ln(2\pi)^n|K_{X^2}|$

$= \frac{1}{2}E_X[\sum_{ij}(\vec{x}_i - \mu_{X_i})^t(\vec{x}_j - \mu_{X_j})(K_{X^2}^{-1})_{ij}] + \frac{1}{2}\ln(2\pi)^n|K_{X^2}|$

$= \frac{1}{2}\sum_{ij} E_X[(\vec{x}_j - \mu_{X_j})^t(\vec{x}_i - \mu_{X_i})](K_{X^2}^{-1})_{ij} + \frac{1}{2}\ln(2\pi)^n|K_{X^2}|$

$= \frac{1}{2}\sum_{ij}[(K_{X^2})_{ji}(K_{X^2}^{-1})_{ij}] + \frac{1}{2}\ln(2\pi)^n|K_{X^2}|$

$= \frac{1}{2}\sum_{j}[(K_{X^2})_{jj}(K_{X^2}^{-1})_{jj}] + \frac{1}{2}\ln(2\pi)^n|K_{X^2}|$

$= \frac{1}{2}\sum_{j} I_{jj} + \frac{1}{2}\ln(2\pi)^n|K_{X^2}|$

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

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

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 sizes k_1 and k_2 respectively, with a multivariate Gaussian distribution:*

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

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:

$h(X_1|X_2) = -\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p_{X_1X_2}(\vec{x_1}, \vec{x_2}) \ln\{p_{X_1|X_2}(\vec{x_1}, \vec{x_2})\} \overrightarrow{dx_1} \overrightarrow{dx_2}$

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

$P_{X_1|X_2}(\vec{x_1}, \vec{x_2}) = (2\pi)^{-\frac{k_1}{2}} |K_{X_1^2|X_2}|^{-\frac{1}{2}} \exp\{-\frac{(\vec{x_1} - \vec{\nu_{X_1/X_2}})^t K_{X_1^2|X_2} (\vec{x_1} - \vec{\nu_{X_1/X_2}})}{2}\}$

So we can write:

$h(X_1|X_2)$

$= -\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p_{X_1X_2}(\vec{x_1}, \vec{x_2}) [-\frac{1}{2}(\vec{x_1} - \nu_{X_1|X_2})^t K_{X^2}^{-1} (\vec{x_1} - \nu_{X_1|X_2}) - \ln(\sqrt{2\pi})^{k_1} |K_{X_1^2|X_2}|^{\frac{1}{2}}] \vec{dx}$

$= \frac{1}{2} E_{X_1X_2}[\sum_{ij} \{(\vec{x_1})_i - \nu_{(X_1|X_2)_i}\}^t (K_{X_1^2|X_2}^{-1})_{ij} \{(\vec{x_1})_j - \nu_{(X_1|X_2)_j}\}] + \frac{1}{2} \ln(2\pi)^{k_1} |K_{X_1^2|X_2}|$

$= \frac{1}{2} E_{X_1X_2}[\sum_{ij} \{(\vec{x_1})_i - \nu_{(X_1|X_2)_i}\}^t \{(\vec{x_1})_j - \nu_{(X_1|X_2)_j}\} (K_{X_1^2|X_2}^{-1})_{ij}] + \frac{1}{2} \ln(2\pi)^{k_1} |K_{X_1^2|X_2}|$

$= \frac{1}{2} \sum_{ij} E_{X_1X_2}[\{(\vec{x_1})_j - \nu_{(X_1|X_2)_j}\}^t \{(\vec{x_1})_i - \nu_{(X_1|X_2)_i}\}] (K_{X_1^2|X_2}^{-1})_{ij} + \frac{1}{2} \ln(2\pi)^{k_1} |K_{X_1^2|X_2}|$

$= \frac{1}{2} \sum_{ij} (K_{X_1^2|X_2})_{ji} (K_{X_1^2|X_2}^{-1})_{ij}) + \frac{1}{2} \ln(2\pi)^{k_1} |K_{X_1^2|X_2}|$

$= \frac{1}{2} \sum_{j} (K_{X_1^2|X_2})_{jj} (K_{X_1^2|X_2}^{-1})_{jj}) + \frac{1}{2} \ln(2\pi)^{k_1} |K_{X_1^2|X_2}|$

$= \frac{1}{2} \sum_{j} I_{jj} + \frac{1}{2} \ln(2\pi)^{k_1} |K_{X_1^2|X_2}|$

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

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

3.4 Mutual information for two Gaussian random vectors

Theorem: *If we consider two Gaussian random vectors $\vec{x}_1$ and $\vec{x}_2$ of sizes k_1 and k_2 respectively, then we can compute the mutual information $I(X_1, X_2)$ as follows:*

$$I(X_1, X_2) = \frac{1}{2} \ln \frac{|K_{X_1^2}|.|K_{X_2^2}|}{|K_{(X_1,X_2)^2}|}$$

Proof:

$I(X_1, X_2)$

$= h(X_1) - h(X_1|X_2)$

$= h(X_1) + h(X_2) - h(X_1, X_2)$

$= \frac{1}{2}\ln(2\pi e)^{k_1}|K_{X_1^2}| + \frac{1}{2}\ln(2\pi e)^{k_2}|K_{X_2^2}| - \frac{1}{2}(2\pi e)^{k_1+k_2}|K_{(X_1X_2)^2}|$

$= \frac{1}{2}\ln \frac{|K_{X_1^2}|.|K_{X_2^2}|}{|K_{(X_1,X_2)^2}|}$

Corollary: *If we consider two Gaussian variables X_1 and X_2, then we can compute the mutual information $I(X_1, X_2)$ as follows:*

$$I(X_1, X_2) = -\frac{1}{2}\ln(1 - \rho^2_{X_1X_2})$$

Proof:

For $k_1 = k_2 = 1$, we can write:

$I(X_1, X_2)$

$= \frac{1}{2}\ln \frac{|K_{X_1^2}|.|K_{X_2^2}|}{|K_{(X_1,X_2)^2}|}$

$= \frac{1}{2}\ln(2\pi e)K_{X_1^2} + \frac{1}{2}\ln(2\pi e)K_{X_2^2} - \frac{1}{2}\ln(2\pi e)^2.\begin{vmatrix} K_{X_1^2} & K_{X_1X_2} \\ K_{X_1X_2} & K_{X_2^2} \end{vmatrix}$

$= \frac{1}{2}\ln(2\pi e)K_{X_1^2} + \frac{1}{2}\ln(2\pi e)K_{X_2^2} - \frac{1}{2}\ln(2\pi e)^2 - \frac{1}{2}\ln(K_{X_1^2}K_{X_2^2} - K^2_{X_1X_2})$

$= \frac{1}{2}\ln(2\pi e)K_{X_1^2} + \frac{1}{2}\ln(2\pi e)K_{X_2^2} - \frac{1}{2}\ln(2\pi e)^2 - \frac{1}{2}\ln K_{X_1^2}.K_{X_2^2}(1 - \rho^2_{X_1X_2})$

$= \frac{1}{2}\ln(2\pi e)^2 K_{X_1^2}K_{X_2^2} - \frac{1}{2}\ln(2\pi e)^2 K_{X_1^2}K_{X_2^2} - \frac{1}{2}\ln(1 - \rho^2_{X_1X_2})$

$= -\frac{1}{2}\ln(1 - \rho^2_{X_1X_2})$

3.5 Conditional mutual information between two Gaussian random vectors given a Gaussian random vector

Theorem: *Given three concatenated random vectors* $(\vec{x}_1, \vec{x}_2, \vec{x}_3)$*, defined on the sets* $\mathbb{X}_1$*,* $\mathbb{X}_2$ *and* $\mathbb{X}_3$ *of sizes n ,m and l respectively, with a multivariate Gaussian distribution* $p_{X_1X_2X_3}(\vec{x_1}, \vec{x_2}, \vec{x_3})$*, we can compute the conditional mutual information between two Gaussian random vectors given a Gaussian random vector as follows:*

$$I(X_1, X_2|X_3) = I(X_1, X_2) + \frac{1}{2}\ln\{(2\pi e)\frac{|K_{X_3^2|X_1}||K_{X_3^2|X_2}|}{|K_{X_3^2}|.|K_{X_3^2|X_1X_2}|}\}$$

Proof:

$I(X_1, X_2|X_3)$

$= h(X_1, X_3) + h(X_2, X_3) - h(X_3) - h(X_1, X_2, X_3)$

$= \frac{1}{2}\ln\{(2\pi)^{k_1+k_3}|K_{(X_1X_3)^2}|\} + \frac{1}{2}\ln\{(2\pi)^{k_2+k_3}|K_{(X_2X_3)^2}|\} - \frac{1}{2}\ln\{(2\pi e)^{k_3}|K_{X_3^2}|\} - \frac{1}{2}\ln\{(2\pi e)^n|K_{(X_1X_2X_3)^2}|\}$

$= \frac{1}{2}\ln\{(2\pi e)^{\frac{k_1+k_2+2k_3}{k_3+n}}\frac{|K_{(X_1,X_3)^2}|.|K_{(X_2X_3)^2}|}{|K_{X_3^2}|.|K_{(X_1X_2X_3)^2}|}\}$

However: $k_3 = n - k_1 - k_2$: $\frac{k_1+k_2+2k_3}{k_3+n} = \frac{k_1+k_2+2(n-k_1-k_2)}{n+n-k1-k_2} = 1$

$I(X_1, X_2|X_3)$

$= \frac{1}{2}\ln\{(2\pi e)\frac{|K_{(X_1,X_3)^2}|.|K_{(X_2X_3)^2}|}{|K_{X_3^2}|.|K_{(X_1X_2X_3)^2}|}\}$

$= \frac{1}{2}\{(2\pi e)\frac{|K_{X_1^2}|.|K_{X_3^2|X_1}|.|K_{X_2^2}|.|K_{X_3^2|X_2}|}{|K_{X_3^2}|.|K_{(X_1X_2)^2}|.|K_{X_3^2|X_1X_2}|}$

$= \frac{1}{2}\ln\frac{|K_{X_1^2}|.|K_{X_2^2}|}{|K_{(X_1,X_2)^2}|} + \frac{1}{2}\ln\{(2\pi e)\frac{|K_{X_3^2|X_1}||K_{X_3^2|X_2}|}{|K_{X_3^2}|.|K_{X_3^2|X_1X_2}|}\}$

$= I(X_1, X_2) + \frac{1}{2}\ln\{(2\pi e)\frac{|K_{X_3^2|X_1}||K_{X_3^2|X_2}|}{|K_{X_3^2}|.|K_{X_3^2|X_1X_2}|}\}$

3.6 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 sizes k_1, k_2 and $k_3 = 1$ respectively, with the multivariate Gaussian distribution $\mathcal{N}(\mu_X, K_{X^2})$ 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 $W_{(X_1,X_2)^2} = K^{-1}_{(X_1,X_2)^2}$ and prove the following relation:

$$W_{X_1^2} = K^{-1}_{X_1^2} + W_{X_1X_2}.W^{-1}_{X_2^2}.W_{X_2X_1}$$

$$K_{X^2}W_{X^2} = \begin{pmatrix} K_{X_1^2}W_{X_1^2} + K_{X_1X_2}W_{X_2X_1} & K_{X_1^2}W_{X_1X_2} + K_{X_1X_2}W_{X_2^2} \\ K_{X_2X_1}W_{X_1^2} + K_{X_2^2}W_{X_2X_1} & K_{X_2X_1}W_{X_1X_2} + K_{X_2^2}W_{X_2^2} \end{pmatrix} = \begin{pmatrix} I_{k_1,k_1} & 0 \\ 0 & I_{k_2,k_2} \end{pmatrix}$$

$K_{X_1X_2}.W_{X_2^2} + K_{X_1^2}.W_{X_1X_2} = 0$

$K^{-1}_{X_1^2}.K_{X_1X_2}.W_{X_2^2} + W_{X_1X_2} = 0$

$K^{-1}_{X_1^2}.K_{X_1X_2} = -W_{X_1X_2}.W^{-1}_{X_2^2}$

$K_{X_1^2}.W_{X_1^2} + K_{X_1X_2}.W_{X_2X_1} = I_{k_1k_1}$

$W_{X_1^2} = K^{-1}_{X_1^2} - K^{-1}_{X_1^2}.K_{X_1X_2}.W_{X_2X_1}$

$W_{X_1^2} = K^{-1}_{X_1^2} + W_{X_1X_2}.W^{-1}_{X_2^2}.W_{X_2X_1}$

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

$(K_{X_3,X_1}, K_{X_3X_2}).K^{-1}_{(X_1X_2)^2}.\begin{pmatrix} K_{X_1X_3} \\ K_{X_2X_3} \end{pmatrix}$

$= (K_{X_3,X_1}, K_{X_3X_2}).W_{(X_1X_2)^2}.\begin{pmatrix} K_{X_1X_3} \\ K_{X_2X_3} \end{pmatrix}$

$= (K_{X_3X_1}, K_{X_3X_2})\begin{pmatrix} W_{X_1^2} & W_{X_1X_2} \\ W_{X_2X_1} & W_{X_2^2} \end{pmatrix}\begin{pmatrix} K_{X_1X_3} \\ K_{X_2X_3} \end{pmatrix}$

$= (K_{X_3X_1})^t W_{X_1^2}(K_{X_1X_3}) + (K_{X_3X_1})W_{X_1X_2}(K_{X_2X_3}) + (K_{X_3X_2})W_{X_2X_1}(K_{X_1X_3})$

$+(K_{X_3X_2}).W_{X_2^2}(K_{X_2X_3})$

Using the relation: $W_{X_1^2} = K^{-1}_{X_1^2} + W_{X_1X_2}.W^{-1}_{X_2^2}.W_{X_2X_1}$:

$= (K_{X_3X_1}).K^{-1}_{X_1^2}.(K_{X_1X_3}) + (K_{X_3X_1}).W_{X_1X_2}.W^{-1}_{X_2^2}.W_{X_2X_1}.(K_{X_1X_3})$

$+(K_{X_3X_1}).W_{X_1X_2}.(K_{X_2X_3}) + (K_{X_3X_2}).W_{X_2X_1}.(K_{X_1X_3})$

$+(K_{X_3X_2}).W_{X_2^2}.(K_{X_2X_3})$

$= [(K_{X_2X_3}) + W_{X_2^2}^{-1}.W_{X_2X_1}.(K_{X_1X_3})]^t.[W_{X_2^2}.(K_{X_2X_3}) + W_{X_2X_1}.(K_{X_1X_3})]$

$+(K_{X_3X_1}).K_{X_1^2}^{-1}.(K_{X_1X_3})$

$= (K_{X_3X_1}) \cdot K_{X_1^2}^{-1}.(K_{X_1X_3})$

$+[(K_{X_2X_3}) + W_{X_2^2}^{-1}.W_{X_2X_1}.(K_{X_1X_3})]^t.W_{X_2^2}.[(K_{X_2X_3}) + W_{X_2^2}^{-1}.W_{X_2X_1}.(K_{X_1X_3})]$

However both following quadratics forms are equivalents :

$$(K_{X_3X_1}, K_{X_3X_2}) \begin{pmatrix} W_{X_1^2} & W_{X_1X_2} \\ W_{X_2X_1} & W_{X_2^2} \end{pmatrix} \begin{pmatrix} K_{X_1X_3} \\ K_{X_2X_3} \end{pmatrix} = (K_{X_3X_2}, K_{X_3X_1}) \begin{pmatrix} W_{X_2^2} & W_{X_2X_1} \\ W_{X_1X_2} & W_{X_1^2} \end{pmatrix} \begin{pmatrix} K_{X_2X_3} \\ K_{X_1X_3} \end{pmatrix}$$

and are positive semidefinite if and only if:

$W_{X_1^2} \geq 0$, $W_{X_2^2} - W_{X_2X_1}.W_{X_1^2}^{-1}.W_{X_1X_2} \geq 0$

but yet

$W_{X_2^2} \geq 0$, $W_{X_1^2} - W_{X_1X_2}.W_{X_2^2}^{-1}.W_{X_2X_1} \geq 0$

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

$$(K_{X_3,X_1}, K_{X_3X_2}) \cdot K_{(X_1X_2)^2}^{-1} \cdot \begin{pmatrix} K_{X_1X_3} \\ K_{X_2X_3} \end{pmatrix} \geq (K_{X_3X_1}).K_{X_1^2}^{-1}.(K_{X_1X_3}) \geq 0$$

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

$$K_{X_3^2} - (K_{X_3,X_1}, K_{X_3X_2}).K_{(X_1X_2)^2}^{-1} \cdot \begin{pmatrix} K_{X_1X_3} \\ K_{X_2X_3} \end{pmatrix} \leq K_{X_3^2} - (K_{X_3X_1}).K_{X_1^2}^{-1}.(K_{X_1X_3}) \leq K_{X_3^2}$$

$K_{X_3^2|X_1X_2} \leq K_{X_3^2|X_1} \leq K_{X_3^2}$

$\frac{1}{2}\ln|K_{X_3^2|X_1X_2}| + \frac{1}{2}\ln(2\pi e)^n \leq \frac{1}{2}\ln|K_{X_3^2|X_1}| + \frac{1}{2}\ln(2\pi e)^n \leq \frac{1}{2}\ln|K_{X_3^2}| + \frac{1}{2}\ln(2\pi e)^n$

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 a set of n nodes X which verify the following proprieties:*

- *For each node, we attribute n random variables included in a random vector $\vec{x} = (x_1, x_2, ..., x_n)$*
- *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:*

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

where we have: $p_{X_j|pa(X_j)}(x_j, Pa(x_j)) = \frac{p_{X_j,Pa(X_j)}(x_j, Pa(x_j))}{p_{Pa(X_j)}(Pa(x_j))}$.

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

4.2 Differential entropy of a Bayesian network

If $\mathcal{B}$ is a Bayesian network to which we attribute a Gaussian random vector $\vec{x} = (X_1, X_2, ..., X_n)$ to the set of Gaussian random variables X, we can compute the differential entropy of this network as follows:

$h(X|\mathcal{B})$

$= -E_X[\ln p_X(\vec{x}|\mathcal{B})]$

$= -E_X[\ln \prod_{x_j \in X} p_{X_j|Pa(X_j)}(x_j, Pa(x_j))]$

$= -E_X[\sum_{x_j \in X} \ln p_{X_j|Pa(X_j)}(x_j, Pa(x_j))]$

$= \sum_{x_j \in X} -E_X[\ln p_{X_j|Pa(X_j)}(x_j, Pa(x_j))]$

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

$= \frac{1}{2} \sum_{x_j \in X} \ln(2\pi e) K_{X_j^2|Pa(X_j)}$

$= \frac{1}{2} \sum_{x_j \in X} \ln(K_{X_j^2|Pa(X_j)}) + \frac{1}{2} \ln(2\pi e)^n$

4.3 Entropy of a chain rule

If $\mathcal{B}^C$ is a a Bayesian network with a chain to which we attribute a Gaussian random vector $\vec{x} = (X_1, X_2, ..., X_n) \sim \mathcal{N}(\mu_X, K_{X^2})$ to the set of Gaussian random variables X, we can compute the differential entropy joint by a chain rule as follows:

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

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.

4.4 Entropy for isolated nodes

If $\mathcal{B}^R$ is a Bayesian network with isolated nodes to which we attribute a Gaussian random vector $\vec{x} = (X_1, X_2, ..., X_n) \sim \mathcal{N}(\mu_X, K_{X^2})$ to the set of Gaussian random variables X, we can compute the differential entropy joint as follows:

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

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

Theorem *Given a Gaussian random vector $\vec{x} = (x_1, x_2, ..., x_n)$ 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:*

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

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

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

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

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

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 $\mathcal{O}(X)$:

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

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):

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

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

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

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

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

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

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

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 $K_{X_j^2|Pa(X_j)})$ 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:

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

$= \frac{1}{2} \ln(\prod_{x_j \in \mathcal{O}(X)} K_{X_j^2|Pa(X_j)}) + \frac{1}{2} \ln(2\pi.e)^n$

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

and for the upper boundary :

$\sum_{X_j \in \mathcal{O}(X)} h(X_j)$

$= \frac{1}{2} \sum_{X_j \in \mathcal{O}(X)} \ln(K_{X_j^2}) + \frac{1}{2} \ln(2\pi e)^n$

$= \frac{1}{2} \sum_{i=1}^{n} \ln(K_{X_{ii}^2}) + \frac{1}{2} \ln(2\pi e)^n$

Finally, we obtain:

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

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 Likelihood function for learning data from Bayesian network

In this section, we will expose the likelihood function to introduce subsequently the scores AIC, BIC and the entropy absorption of Bayesian network.

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

As we have the relation:

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

We can put:

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

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

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

$$P_{X_j|Pa(X_j)}(x_j, Pa(x_j)) = (2\pi)^{-\frac{1}{2}} K^{-\frac{1}{2}}_{X_j^2|Pa(X_j)} \exp\{-\frac{(x_j-\beta_{(X_j,Pa(X_j))}.Pa(x_j)-\beta_{X_j})^2}{2.K_{X_j^2|Pa(X_j)}}\}$$

For N points of a continuous data matrix, we then compute the likelihood function:

$$L(D|\beta_{X_j},\beta_{(X_j,Pa(X_j))},K_{X_j^2|Pa(X_j)})$$

$$= \ln \prod_{i=1}^{N} \prod_{X_j \in X} P_{X_j|Pa(X_j)}(x_j, Pa(x_j))$$

$$= \ln \prod_{i=1}^{N} \prod_{X_j \in X} (2\pi)^{-\frac{1}{2}} K^{-\frac{1}{2}}_{X_j^2|Pa(X_j)} \exp\{-\frac{(x_{ij} - \beta_{(X_j,Pa(X_j))}.Pa(x_{ij}) - \beta_{X_j})^2}{2.K_{X_j^2|Pa(X_j)}}\}$$

$$= \sum_{X_j \in X} \sum_{i=1}^{N} -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(K_{X_j^2|Pa(X_j)}) - \frac{1}{2K_{X_j^2|Pa(X_j)}}.(x_{ij} - \beta_{(X_j,Pa(X_j))}.Pa(x_{ij}) - \beta_{X_j})^2$$

$$= \sum_{X_j \in X} -\frac{N}{2}\ln(2\pi) - \frac{N}{2}.\ln(K_{X_j^2|Pa(X_j)}) - \frac{1}{2K_{X_j^2|Pa(X_j)}}.\sum_{i=1}^{N}(x_{ij} - \beta_{(X_j,Pa(X_j))}.Pa(x_{ij}) - \beta_{X_j})^2$$

$$= \sum_{X_j \in X} \frac{-N}{2}\ln(2\pi.K_{X_j^2|Pa(X_j)}) - \frac{1}{2K_{X_j^2|Pa(X_j)}}.\sum_{i=1}^{N}(x_{ij} - \beta_{(X_j,Pa(X_j))}.Pa(x_{ij}) - \beta_{X_j})^2$$

6 AIC, AICc and BIC Gaussian score

From the likelihood function that was described in the previous section, we will now expose the AIC, AICc and BIC scores used to learn continuous data from Bayesian networks.

If we put the residual value:

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

However: $\frac{SSR(X_j|Pa(X_j))}{2K_{X_j^2|Pa(X_j)}} = \frac{N-\#Pa(X_j)-1}{2}$:

$$L(D|\beta_{X_j},\beta_{(X_j,Pa(X_j))},K_{X_j^2|Pa(X_j)}) = \sum_{X_j \in X} \{ \frac{-N}{2} \ln(2\pi.K_{X_j^2|Pa(X_j)}) - \frac{(N-\#Pa(X_j)-1)}{2} \}$$

$$L(D|\beta_{X_j},\beta_{(X_j,Pa(X_j))},K_{X_j^2|Pa(X_j)}) = \sum_{X_j \in X} \frac{-N}{2} \ln(2\pi.K_{X_j^2|Pa(X_j)}) - \sum_{X_j \in X} \frac{(N-\#Pa(X_j)-1)}{2}$$

$$L(D|\beta_{X_j},\beta_{(X_j,Pa(X_j))},K_{X_j^2|Pa(X_j)}) = -N \sum_{X_j \in X} \frac{1}{2} \ln(2\pi.K_{X_j^2|Pa(X_j)}) - \frac{(n.N-q-n)}{2}$$

$$L(D|\beta_{X_j},\beta_{(X_j,Pa(X_j))},K_{X_j^2|Pa(X_j)}) = -N \sum_{X_j \in X} \frac{1}{2} \ln(2\pi.K_{X_j^2|Pa(X_j)}) - \frac{(n.(N-1)-q)}{2}$$

We can build the AIC, AICc (AIC with a corrector for a small data matrix) and the BIC score:

$$AIC(D|\mathcal{B}) = \overline{L}(D|\beta_{X_j},\beta_{(X_j,Pa(X_j))},K_{X_j^2|Pa(X_j)}) + (q+2n)$$

$$AICc(D|\mathcal{B}) = AIC(D|\mathcal{B}) + \frac{(q+2n).(q+2n+1)}{N-q-2n-1}$$

$$BIC(D|\mathcal{B}) = \overline{L}(D|\beta_{X_j},\beta_{(X_j,Pa(X_j))},K_{X_j^2|Pa(X_j)}) + \frac{(q+2n)\ln(N)}{2}$$

7 Differential entropy absorption of a Bayesian network

7.1 Bayesian Network differential entropy absorption and Kullback-leibler divergence

By using the inequalities:

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

We can bound the entropy variation $\sum_{X_j \in X} h(X_j|Pa(X_j)) - h(X_1, X_2, ..., X_n)$:

$$0 \leq \sum_{X_j \in X} h(X_j|Pa(X_j)) - h(X_1, X_2, ..., X_n) \leq \sum_{X_j \in X} h(X_j) - h(X_1, X_2, ..., X_n)$$

The entropy variation is the Kullback-leibler divergence:

$$\sum_{X_j \in X} h(X_j|Pa(X_j)) - h(X_1, X_2, ..., X_n) = -E_X[\ln\{\frac{p_X(\vec{x}|\mathcal{B})}{p_X(\vec{x})}\}] = D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B}))$$

The Kullback-leibler divergence satisfies the inequality:

$$0 \leq D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B})) \leq D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B}^R))$$

where we have: $D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B}^R)) = \sum_{X_j \in X} h(X_j) - h(X_1, X_2, ..., X_n)$

We can express the entropy of the Bayesian network $h(X|\mathcal{B})$ as a function of the Kullback-leibler divergence as follows:

$$h(X|\mathcal{B}) = h(X_1, X_2, ..., X_n) + D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B}))$$

The Bayesian network correctly absorbs the entropy of the data matrix $h(X_1, X_2, ..., X_n)$ when the Kullback-leibler divergence $D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B}))$ is close to 0. If the Kullback-leibler divergence approaches 0 then the likelihood between the data matrix and the Bayesian network increases more and more. In the next section, this will be expressed, from a theorem relating the loss of likelihood and the Kullback-leibler divergence.

7.2 Kullback-leibler divergence and loss of likelihood

Theorem *If $\mathcal{B}$ is a Bayesian network then the Kullback-leibler divergence is related to the loss of likelihood $\frac{\Delta\overline{L}}{\overline{L}_{min}}$ by the following relation:*

$$D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B})) = \frac{\Delta\overline{L}}{\overline{L}_{min}}.\{h(X_1, X_2, ..., X_n) + \frac{n(N-2)}{2N}\}$$

where we have:

$\overline{L}(D|\mathcal{B}) = N.\sum_{X_j \in X} h(X_j|Pa(X_j)) + \frac{n.(N-2)}{2}$

$\overline{L}_{min} = \frac{N}{2}\ln|K_{X^2}| + \frac{N}{2}\ln(2\pi e)^n + \frac{n.(N-2)}{2}$

$\frac{\Delta\overline{L}}{\overline{L}_{min}} = \frac{\overline{L}(D|\mathcal{B}) - \overline{L}_{min}}{\overline{L}_{min}}$

$\frac{\overline{L}(D|\mathcal{B})}{\overline{L}_{min}} = 1 + \frac{\Delta\overline{L}}{\overline{L}_{min}}$

Proof:

Consider the following likelihood function:

$\overline{L}(D|\beta_{X_j}, \beta_{(X_j, Pa(X_j))}, K_{X_j^2|Pa(X_j)})$

$= \sum_{X_j \in X} \frac{N}{2}\ln(2\pi.K_{X_j^2|Pa(X_j)}) + \frac{1}{2K_{X_j^2|Pa(X_j)}}.\sum_{i=1}^{N}(x_{ij} - \beta_{(X_j, Pa(X_j))}.Pa(x_{ij}) - \beta_{X_j})^2$

Using $\frac{SSR(X_j|Pa(X_j))}{2K_{X_j^2|Pa(X_j)}} = \frac{N-1}{2}$ and $h(X_j|Pa(X_j)) = \frac{1}{2}\ln(2\pi e K_{X_j^2|Pa(X_j)})$, we can write:

$\overline{L}(D|\beta_{X_j}, \beta_{(X_j, Pa(X_j))}, K_{X_j^2|Pa(X_j)}) = N.\sum_{X_j \in X} h(X_j|Pa(X_j)) + \frac{n.(N-2)}{2}$

With the bounds, we obtain the inequalities:

$\frac{N}{2}\ln|K_{X^2}| + \frac{N}{2}\ln(2\pi e)^n + \frac{n.(N-2)}{2}$

$\leq \overline{L}(D|\beta_{X_j}, \beta_{(X_j, Pa(X_j))}, K_{X_j^2|Pa(X_j)})$

$\leq \frac{N}{2}\sum_{i=1}^{n}\ln(K_{X_{ii}^2}) + \frac{N}{2}\ln(2\pi e)^n + \frac{n.(N-2)}{2}$

we can also write:

$N.h(X_1, X_2, ..., X_n) + \frac{n.(N-2)}{2}$

$\leq N.\sum_{X_j \in X} h(X_j|Pa(X_j)) + \frac{n.(N-2)}{2}$

$\leq N.\sum_{X_j \in X} h(X_j) + \frac{n.(N-2)}{2}$

If we put:

$\overline{L}_{min} = N.h(X_1, X_2, ..., X_n) + \frac{n.(N-2)}{2}$,

$\overline{L}_{max} = N.\sum_{X_j \in X} h(X_j) + \frac{n.(N-2)}{2}$,

$\overline{L}(D|\mathcal{B}) = \overline{L}(D|\beta_{X_j}, \beta_{(X_j, Pa(X_j))}, K_{X_j^2|Pa(X_j)})$

and $D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B})) = \sum_{X_j \in X} h(X_j|Pa(X_j)) - h(X_1, X_2, ..., X_n)$,

we can make the relation between the Kullback-leibler $D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B}))$ and the loss of likelihood $\frac{\overline{L}(D|\mathcal{B}) - \overline{L}_{min}}{\overline{L}_{min}}$:

$$\frac{\overline{L}(D|\mathcal{B}) - \overline{L}_{min}}{\overline{L}_{min}} = \frac{\Delta\overline{L}}{\overline{L}_{min}} = \frac{N.D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B}))}{N.h(X_1, X_2, ..., X_n) + n.\frac{(N-2)}{2}}$$

Finally we obtain:

$$D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B})) = \frac{\Delta\overline{L}}{\overline{L}_{min}}.\{h(X_1, X_2, ..., X_n) + \frac{n(N-2)}{2N}\}$$

where the loss of likelihood satisfies the equality:

$$\frac{\overline{L}(D|\mathcal{B})}{\overline{L}_{min}} = 1 + \frac{\Delta\overline{L}}{\overline{L}_{min}}$$

Note that the bigger loss of likelihood is:

$$\frac{\Delta\overline{L}_{max}}{\overline{L}_{min}} = \frac{\overline{L}_{max} - \overline{L}_{min}}{\overline{L}_{min}} = \frac{N.\{\sum_{X_j \in X} h(X_j) - h(X_1, X_2, ..., X_n)\}}{N.h(X_1, X_2, ..., X_n) - \frac{n(N-2)}{2}} = \frac{N.D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B}^R))}{N.h(X_1, X_2, ..., X_n) - \frac{n(N-2)}{2}}$$

7.3 Learning continuous data matrix algorithm based on the entropy absorption of Bayesian network

When learning the data matrix, we will start by presenting the bigger loss of likelihood $\frac{\Delta\overline{L}_{max}}{\overline{L}_{min}}$ that we can obtain:

$$0 \leq \frac{\Delta\overline{L}}{\overline{L}_{min}} = \frac{\overline{L}(D|\mathcal{B}) - \overline{L}_{min}}{\overline{L}_{min}} \leq \frac{\Delta\overline{L}_{max}}{\overline{L}_{min}} = \frac{\overline{L}_{max} - \overline{L}_{min}}{\overline{L}_{min}}$$

and the bigger Kullback-leibler divergence:

$$D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B}^R)) = \sum_{X_j \in X} h(X_j) - h(X_1, X_2, ..., X_n)$$

then we will set a loss of likelihood value λ_{max} not to be exceeded:

$$\frac{\overline{L}(D|\mathcal{B})}{\overline{L}_{min}} = 1 + \frac{\Delta\overline{L}}{\overline{L}_{min}} \leq 1 + \lambda_{max}$$

This limit loss of likelihood value will give us an upper bound on the Kullback-leibler divergence not to be exceeded:

$$D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B})) \leq \lambda_{max}.\{h(X_1, X_2, ..., X_n) + \frac{n(N-2)}{2N}\}$$

The algorithm starts with a Bayesian chain network for a fixed topological order. The number of chain networks having n nodes linked to a topological order is $n!$, we must therefore consider as many chain networks as possible at the start. For each chain network, we must then iteratively remove the nodes causing the smallest conditional entropy variation to get the smallest Kullback-leibler divergence variation verifying the previous inequality. The smallest variation of conditional entropy caused by the removal of nodes allows to remove the nodes causing the weakest conditionings and to keep the nodes causing the higher conditionings. Among all the candidate Bayesian networks, we will choose the Bayesian network having both the smallest number of edges and the smallest Kullback-leibler divergence to obtain the best likelihood between the data matrix and the Bayesian network. In this report we will consider only one topological order and not forget that we have to apply this method to many topological orders.

7.4 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):

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

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

8 Learning a continuous data matrix from Bayesian network Example

We consider the continuous data matrix and the topological sort $(X_5, X_2, X_4, X_3, X_1, X_6)$:

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	3.460038	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

In this example, we consider the topological sort $(X_5, X_2, X_4, X_3, X_1, X_6)$. The bigger loss of likelihood is:

$$\frac{\Delta \overline{L}_{max}}{\overline{L}_{min}} = \frac{\overline{L}_{max} - \overline{L}_{min}}{\overline{L}_{min}} = 0.1540101$$

and the bigger Kullback-leibler divergence $D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B}^R))$ is :

$$D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B}^R)) = \frac{\Delta \overline{L}_{max}}{\overline{L}_{min}}.\{h(X_1, X_2, ..., X_n) + \frac{n(N-2)}{2N}\} = 2.95544$$

We will fixed the maximum loss of likelihood to $\lambda_{max} = 2.10^{-3}$. Using conditional entropy, we will remove the weakest conditionings and keep the higher conditionings. The Kullback-leibler divergence will have to verify the inequality:

$$D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B})) \leq \lambda_{max}.\{h(X_1, X_2, ..., X_n) + \frac{n(N-2)}{2N}\} = 0.03837982$$

Note: We use the notation $\tilde{X_j}$ to remove the node X_j.

The algorithm start with the following chain rule:

$h(X_5)+h(X_2|X_5)+h(X_4|X_5X_2)+h(X_3|X_5X_2X_4)+h(X_1|X_5X_2X_4X_3)+h(X_6|X_5X_2X_4X_3X_1)$

= 1.404689 + 5.477484 + 0.3532928 + 5.06253 + 2.359493 + 1.59242

= 16.24991

$$D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B})) = 0 < 0.03837982$$

$h(X_5)+h(X_2|X_5)+h(X_4|X_5X_2)+h(X_3|X_5X_2X_4)+h(X_1|X_5X_2X_4X_3)+h(X_6|X_5\tilde{X_2}X_4X_3X_1)$

= 1.404689 + 5.477484 + 0.3532928 + 5.06253 + 2.359493 + 1.592658

= 16.25015

$$D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B})) = 0.00024 < 0.03837982$$

$h(X_5)+h(X_2|X_5)+h(X_4|X_5X_2)+h(X_3|X_5X_2X_4)+h(X_1|X_5X_2\tilde{X_4}X_3)+h(X_6|X_5\tilde{X_2}X_4X_3X_1)$

= 1.404689 + 5.477484 + 0.3532928 + 5.06253 + 2.361073 + 1.592658

= 16.25173

$$D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B})) = 0.00182 < 0.03837982$$

$h(X_5)+h(X_2|X_5)+h(X_4|X_5X_2)+h(X_3|\tilde{X_5}X_2X_4)+h(X_1|X_5X_2\tilde{X_4}X_3)+h(X_6|X_5\tilde{X_2}X_4X_3X_1)$

= 1.404689 + 5.477484 + 0.3532928 + 5.065635 + 2.361073 + 1.592658

= 16.25483

$$D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B})) = 0.00492 < 0.03837982$$

$$h(X_5)+h(X_2|X_5)+h(X_4|X_5X_2)+h(X_3|\tilde{X}_5X_2\tilde{X}_4)+h(X_1|X_5X_2\tilde{X}_4X_3)+h(X_6|X_5\tilde{X}_2X_4X_3X_1)$$
$$= 1.404689 + 5.477484 + 0.3532928 + 5.067788 + 2.361073 + 1.592658$$
$$= 16.25698$$

$$D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B})) = 0.007069999 < 0.03837982$$

$$h(X_5)+h(X_2|X_5)+h(X_4|X_5X_2)+h(X_3|\tilde{X}_5X_2\tilde{X}_4)+h(X_1|X_5X_2\tilde{X}_4X_3)+h(X_6|X_5\tilde{X}_2\tilde{X}_4X_3X_1)$$
$$= 1.404689 + 5.477484 + 0.3532928 + 5.067788 + 2.361073 + 1.59831$$
$$= 16.26264$$

$$D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B})) = 0.01273 < 0.03837982$$

$$h(X_5)+h(X_2|X_5)+h(X_4|X_5X_2)+h(X_3|\tilde{X}_5X_2\tilde{X}_4)+h(X_1|X_5\tilde{X}_2\tilde{X}_4X_3)+h(X_6|X_5\tilde{X}_2\tilde{X}_4X_3X_1)$$
$$= 1.404689 + 5.477484 + 0.3532928 + 5.067788 + 2.366194 + 1.59831$$
$$= 16.26776$$

$$D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B})) = 0.01785 < 0.03837982$$

$$h(X_5)+h(X_2|X_5)+h(X_4|X_5\tilde{X}_2)+h(X_3|\tilde{X}_5X_2\tilde{X}_4)+h(X_1|X_5\tilde{X}_2\tilde{X}_4X_3)+h(X_6|X_5\tilde{X}_2\tilde{X}_4X_3X_1)$$
$$= 1.404689 + 5.477484 + 0.3616165 + 5.067788 + 2.366194 + 1.59831$$
$$= 16.27608$$

$$D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B})) = 0.02617 < 0.03837982$$

$$h(X_5)+h(X_2|X_5)+h(X_4|X_5\tilde{X}_2)+h(X_3|\tilde{X}_5X_2\tilde{X}_4)+h(X_1|X_5\tilde{X}_2\tilde{X}_4X_3)+h(X_6|X_5\tilde{X}_2\tilde{X}_4X_3\tilde{X}_1)$$
$$= 1.404689 + 5.477484 + 0.3616165 + 5.067788 + 2.366194 + 1.605432$$
$$= 16.2832$$

$$D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B})) = 0.03329 < 0.03837982$$

$$h(X_5)+h(X_2|X_5)+h(X_4|X_5\tilde{X}_2)+h(X_3|\tilde{X}_5X_2\tilde{X}_4)+h(X_1|X_5\tilde{X}_2\tilde{X}_4\tilde{X}_3)+h(X_6|X_5\tilde{X}_2\tilde{X}_4X_3\tilde{X}_1)$$
$$= 1.404689 + 5.477484 + 0.3616165 + 5.067788 + 2.488241 + 1.605432$$
$$= 16.40525$$

$$D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B})) = 0.15534 > 0.03837982$$

For the last Bayesian network, the value of the Kullback-leibler divergence exceeds 0.03837982, so we do not consider this case. Finally we get the Bayesian network:

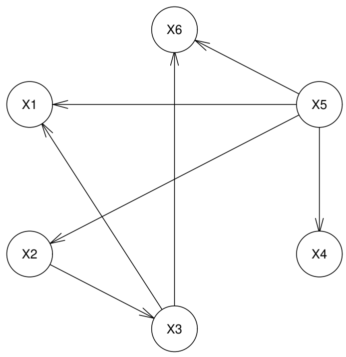

Figure 1: Bayesian network

$h(X_5)+h(X_2|X_5)+h(X_4|X_5\tilde{X_2})+h(X_3|\tilde{X_5}X_2\tilde{X_4})+h(X_1|X_5\tilde{X_2}\tilde{X_4}X_3)+h(X_6|X_5\tilde{X_2}\tilde{X_4}X_3\tilde{X_1})$

$= 1.404689 + 5.477484 + 0.3616165 + 5.067788 + 2.366194 + 1.605432$

$= 16.2832$

$$D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B})) = \sum_{X_j \in X} h(X_j|Pa(X_j)) - h(X_1, X_2, ..., X_n) = 0.03329$$

We can express the entropy of the Bayesian network $h(X|\mathcal{B})$ as a function of the Kullback-leibler divergence as follows:

$$h(X|\mathcal{B}) = h(X_1, X_2, ..., X_n) + D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B})) = 16.24991 + 0.03329 = 16.2832$$

The loss of likelihood $\frac{\Delta\overline{L}}{\overline{L}_{min}}$ is equal to:

$$\frac{\Delta\overline{L}}{\overline{L}_{min}} = \frac{N.D_{KL}(p_X(\vec{x})||p_X(\vec{x}|\mathcal{B}))}{N.h(X_1, X_2, ..., X_n) + n.\frac{(N-2)}{2}} = 1.734766.10^{-3}$$

$$\frac{\overline{L}(D|\mathcal{B})}{\overline{L}_{min}} = 1 + \frac{\Delta\overline{L}}{\overline{L}_{min}} = 1 + 1.734766.10^{-3} = 1.001734766$$

9 Conclusion

In this article, we proposed a data learning algorithm based on the differential entropy absorption of a Bayesian network linked to the loss of likelihood. We exposed a continuous data matrix on which we have detailed the step-by-step learning mechanism.

10 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 :*

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

References

[1] Thomas M Cover and Joy A Thomas. *Elements of information theory*. 2001.

[2] Morris H. DeGroot. *Optimal Statistical Decisions*. John Wiley & Sons, 2005.

[3] Dan Geiger and David Heckerman. Learning gaussian networks. In *Uncertainty Proceedings 1994*, pages 235–243. Elsevier, 1994.

[4] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

[5] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[6] Marco Scutari. Learning bayesian networks with the bnlearn r package. *arXiv preprint arXiv:0908.3817*, 2009.