POLITECNICO DI TORINO

Corso di Laurea in Ingegneria Matematica

Tesi di Laurea Magistrale

Exploring association of several variables using mutual information

Relatori
Prof. Mauro Gasparini
Dr. Pavel Mozgunov

Candidato
Alessandra Serra

A. A. 2017-2018

"*Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone - as the first step"*

Tukey, 1977

Abstract

This work focuses on methods of data exploration using the mutual information and other related information measures. In particular, the author proposes a method to discover pairwise correlations among variables and to classify them into clusters.

The master thesis presents the work done by the author during her internship in Tetra Pak. The core products of the company are the filling machines. The performance of a machine which fills shelf-stable food packages is called the *aseptic performance* and is defined by the long-run ratio between the number of not commercially sterile packages and the total number of packages filled by the machine.
Nowadays, Tetra Pak collects a large amount of data in order to improve the aseptic performance.

The dependencies among variables, in real-world applications like the aseptic performance characterisation, are often unknown and they are almost always characterized by nonlinear relationships. The aim of this study was to find a way to discover correlations among continuous and categorical variables in large datasets. A popular statistic in data mining is a *measure of dependence*. In order to deal with a large amount of variables that could have non-linear dependencies, an adequate measure is required.

The idea is to explore datasets with both continuous and categorical variables and to group them into clusters using a distance based on the *mutual information*. This measure of dependence is well-established in information theory and it can be used to have a better understanding of the relationships among the features.

The principal use of the proposed method is to find a set of uncorrelated variables in order to build predictive models and explain variables of interest.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Tetra Pak

Tetra Pak is the global leader in food processing and packaging solutions. The founder was Dr. Ruben Rausing on the Erik Wallenberg's idea of applying the tetrahedral form to packaging.

The company delivers end-to-end solutions in order to meet the needs of hundreds of millions of consumers in more than 190 countries every day. Tetra Pak provides complete solutions for the processing, packaging and distribution of food products. Dairy products, beverages, ice cream, cheese and prepared food are examples of products that can be processed or packaged in Tetra Pak processing and packaging lines.

Figure 1.1. Tetra Pak logo

The company offers a variety of package shapes, in order to meet customers' requirements and also different packaging materials in order to obtain the best possible

performance with every different food product.

Figure 1.2. Package Portfolio

In addition to a large range of different packages, the company offers a big quantity of filling machines, that are the connection point between packaging material and food product.

Figure 1.3. Tetra Pak A3/Compact Flex

1.2 Aseptic Performance Support

Within Tetra Pak, the Aseptic Performance Support (APS) team provides support, services and trainings to customers in order to develop competences and good production

performances.
The central APS office is based in Modena and the author worked in this team during her internship.

APS team also develops methods and tools for quality data analysis. In addition, the team gives international support to customers, organizing trainings and offering methods to satisfy their quality needs.

1.3 The objective of the work

Nowadays, companies are collecting a large amount of data. We are in the era of Big Data and the volume of available data is growing exponentially. However most of the time there is a lack of knowledge in terms of ability to understand and explore data in an appropriate way. This lack of knowledge and resources could bring economic costs and delays. For example, it can be difficult to create predictive models or build up advanced statistics models without an understanding about the behaviour of the variables. Additional challenges arise as the big quantity of data available contains a lot of missing values, incorrect information and other similar problems.

A part of the author's work in Tetra Pak was to find solutions to these problems by investigating methods for data exploration. Specifically, during the internship, the author worked with different types of datasets concerning process parameters and quality, such as measurements from filling machines and information about packaging material.

All these datasets have some common characteristics:

- the number of variables is high;
- the majority of variables is categorical and only few are on a continuous scale;
- the distributions of the variables are most of the time unknown.

The idea is to explore this type of datasets by grouping the variables into clusters in order to have a better idea of the relationships among the features.

The aim of this work is:

- to compute pairwise correlations among all variables using a measure of dependence;
- to compute hierarchical clustering among variables;
- to find, for each cluster, the variable that shares the most information with Y, given a variable of interest Y.

Chapter 2

Measures of dependence

2.1 Information Theory

Which measure of dependence should be used in order to detect dependencies among categorical and continuous variables?

The usual correlation coefficient, the Pearson coefficient ρ, is commonly used to detect linear associations among continuous variables. Dealing with both categorical and continuous variables other measures are required. We search for measures of dependence and variability designed for general random variables, which work well especially for categorical variables. The information theory approach is used to answer the question.

2.2 Definitions

One of the first information measure was proposed by Shannon (1948) [1] to describe the quantity of information produced by a source. The first measure in the original work was the *entropy*, described as the number of binary digits required to encode a message. Considering a categorical random variable X, entropy $H(X)$ is the amount of information required, on average, to describe X. The second notion of information was the *mutual information*. The mutual information is a measure of a statistical dependence between two sets of random variables. Denoting by X and Y two categorical random variables, the mutual information is the amount of information shared between X and Y.

Definitions concerning the information measures are given below. See, for example, [2], [3].

Definition 2.1. Let X be a categorical random variable with probability density $p_X(x)$. The Shannon entropy of X is defined by

$$H(X) = -\sum_{x\in\chi} p_X(x)\log p_X(x) \tag{2.1}$$

where χ is the support set of the random variable X.

The entropy could also be written as:

$$H(X) = \mathbb{E}\left[\log\frac{1}{p_X(x)}\right]$$

The definition can be extended to a pair of random variables.

Definition 2.2. Let X, Y be two categorical random variables with joint probability density $p_{X,Y}(x,y)$. The joint entropy is defined by

$$H(X,Y) = -\sum_{x\in\chi}\sum_{y\in\gamma} p_{X,Y}(x,y)\log p_{X,Y}(x,y) \tag{2.2}$$

where χ is the support set of the random variable X and γ is the support set of the random variable Y.

Definition 2.3. Let X, Y two categorical random variables with joint probability density $p_{X,Y}(x,y)$ and marginal densities $p_X(x)$ and $p_Y(y)$.
The mutual information between X and Y is defined as:

$$I(X,Y) = \sum_{x\in\chi}\sum_{y\in\gamma} p_{X,Y}(x,y)\log\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} \tag{2.3}$$

with convention $0\log 0 = 0$.

The mutual information could be written as

$$I(X,Y) = \mathbb{E}\left[\log\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right]$$

or in terms of the entropies

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{2.4}$$

The mutual information can also be written using the conditional entropy.

Definition 2.4. Let X, Y be two categorical random variables with joint probability density $p_{X,Y}(x,y)$. The conditional entropy of X given Y is defined by

$$H(X|Y) = -\sum_{x\in\chi}\sum_{y\in\gamma} p_{X,Y}(x,y)\log p_{X|Y}(x|y) \tag{2.5}$$

where

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

It follows that

$$I(X,Y) = \sum_{x\in\chi}\sum_{y\in\gamma} p_{X,Y}(x,y)\log\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} = \sum_{x\in\chi}\sum_{y\in\gamma} p_{X,Y}(x,y)\log\frac{p_{X|Y}(x|y)}{p_X(x)} = \sum_{x\in\chi}\sum_{y\in\gamma} p_{X,Y}(x,y)\log\frac{p_{Y|X}(y|x)}{p_Y(y)}$$

Therefore

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).$$

The relationships among the information measures are given in Figure 2.1

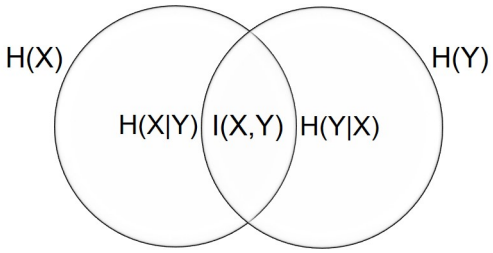


Figure 2.1. Relationships between information measures.

The mutual information has the following properties:

- Symmetry: $I(X,Y) = I(Y,X)$;
- Non negativity: $I(X,Y) \geq 0$;
- $I(X,Y) = 0 \Leftrightarrow X$ and Y are independent.

The natural extension of the finite entropy was introduced by Shannon [1], replacing the sum in Equation (2.1) with the integral.

Definition 2.5. Let X be a continuous random variable with probability density $f(x)$. The differential entropy $h(X)$ of X is defined by

$$h(X) = -\int_\Omega f(x)\log f(x)\mathrm{d}x \tag{2.6}$$

where Ω is the support set of the random variable.

Similarly, the mutual information between two continuous random variables X and Y can be defined as

$$I(X,Y) = \int_{x\in\Omega}\int_{y\in\xi} p_{X,Y}(x,y)\log\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\mathrm{d}x\mathrm{d}y \tag{2.7}$$

or in terms of the entropies as

$$I(X,Y) = h(X) + h(Y) - h(X,Y) \tag{2.8}$$

2.2.1 Properties of entropy and mutual information

Let analyse the main properties of entropy and mutual information, separately for the categorical and continuous cases.

Properties of the entropy and the mutual information of categorical variables
Let X and Y be two categorical random variables. It holds [4]:

a) $H(X) \geq 0$.

   In particular, $H(X) = 0 \Leftrightarrow$ for some i, $p_X(x_i)$ or $p_X(x_j) = 0\ \forall j \neq i$. It means that X is a degenerate random variable;

b) $H(X) \leq \log|A|$ where $|A|$ is the cardinality of the support of X. Equality holds if and only if X has a uniform distribution over A;

c) $H(X) = I(X,X)$;

d) $\max(H(X),H(Y)) \leq H(X,Y) \leq H(X) + H(Y)$;

e) $\min(H(X),H(Y)) \geq I(X,Y) = H(X) - H(X|Y)$.

The differential entropy does not share the same properties.

Properties of the entropy and the mutual information of continuous variables
Let X and Y be two continuous random variables.

a) $h(X)$ can be negative;

b) since $2^{h(X)}$ is the volume of the support set of the random variable X (see [3]), if $h(X) \to \infty$ the support set of the random variable is high and the variable is widely dispersed;

c) $I(X,Y)$ is not bounded;

d) $I(X,Y)$ is invariant under linear transformations, while $h(X)$ is not. In particular [3]:

- $h(X+c) = h(X)$;
- $h(aX) = h(X) + \log|a|$, $a \in \mathbb{R}$;
- $h(AX) = h(X) + \log|det(A)|$, $A \in \mathbb{R}^{n\times n}$.

Consider $X = aZ_1 + b$ and $Y = cZ_2 + d$ where $a,b,c,d \in \mathbb{R}$ and X,Y,Z_1,Z_2 are random variables. In matrix equation

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} a & 0 \\ 0 & c \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} + \begin{pmatrix} b \\ d \end{pmatrix}$$

Denoting matrix $\begin{pmatrix} a & 0 \\ 0 & c \end{pmatrix}$ by A, it follows that

$$\begin{aligned} I(X,Y) &= h(X) + h(Y) - h(X,Y) \\ &= h(Z_1) + \log|a| + h(Z_2) + \log|c| - h(Z_1,Z_2) - \log|det(A)| \\ &= h(Z_1) + \log|a| + h(Z_2) + \log|c| - h(Z_1,Z_2) - \log|ac| \\ &= h(Z_1) + \log|a| + h(Z_2) + \log|c| - h(Z_1,Z_2) - \log|a| - \log|c| \\ &= h(Z_1) + h(Z_2) - h(Z_1,Z_2) \\ &= I(Z_1,Z_2) \end{aligned}$$

The mutual information could be determined analytically in some particular cases, when the joint distribution is given. Moreover, in some special cases, it has been proved that the correlation function and mutual information are directly connected to each other.

In the following sections let focus on two special cases. Firstly, the mutual information between two particular binary random variables is computed. Secondly, we consider two random variables with a Gaussian joint distribution.

2.3 Mutual information of bivariate binary random vector

Let X and Y be Bernoulli with the same marginal distributions.

$$X \sim Bernoulli(p), \qquad Y \sim Bernoulli(p)$$

Let denote with $\pi = \Pr(X = 1, Y = 1)$, $\sigma_X = \sqrt{\mathrm{Var}(X)}$ and the joint distribution be defined as

	$Y=0$	$Y=1$	total
$X=0$	$1-2p+\pi$	$p-\pi$	$1-p$
$X=1$	$p-\pi$	π	p
total	$1-p$	p	1

Table 2.1. Joint distribution of X and Y.

Let compute, firstly, the Pearson correlation coefficient. Since X and Y have the same distribution, the correlation coefficient takes the form

$$\begin{aligned}\rho(X,Y) &= \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X^2} \\ &= \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sigma_X^2} = \frac{\pi - p^2}{p(1-p)}\end{aligned} \tag{2.9}$$

From Equation (2.9), the following equalities can be proven

- $1-\rho(X,Y) = 1 - \dfrac{\pi-p^2}{p(1-p)} = \dfrac{p(1-p)-\pi+p^2}{p(1-p)} = \dfrac{p-\pi}{p(1-p)}$
- $(1-\rho)^2 = \dfrac{p^2+\pi^2-2p\pi}{p^2(1-p)^2} = \dfrac{1}{(1-p)^2} + \dfrac{\pi^2-2p\pi}{p^2(1-p)^2}$
- $\dfrac{\pi^2-2p\pi}{p^2(1-p)^2} = (1-\rho)^2 - \dfrac{1}{(1-p)^2}$

The mutual information can be computed as

$$\begin{aligned}I(X,Y) &= \sum_{x\in\chi}\sum_{y\in\gamma} f_{X,Y}(x,y)\log\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)} \\ &= f_{X,Y}(0,0)\log\frac{f_{X,Y}(0,0)}{f_X(0)f_Y(0)} + 2f_{X,Y}(0,1)\log\frac{f_{X,Y}(0,1)}{f_X(0)f_Y(1)} + f_{X,Y}(1,1)\log\frac{f_{X,Y}(1,1)}{f_X(1)f_Y(1)} \\ &= (1-2p+\pi)\log\left(\frac{1-2p+\pi}{(1-p)^2}\right) + 2(p-\pi)\log\left(\frac{p-\pi}{p(1-p)}\right) + \pi\log\left(\frac{\pi}{p^2}\right)\end{aligned}$$

Using the first equality we can replace $\frac{p-\pi}{p(1-p)}$ by $1-\rho(X,Y)$ and we obtain

$$\begin{aligned}
I(X,Y) &= (1-2p)\log\left(\frac{1-2p+\pi}{(1-p)^2}\right)+\pi\left[\log\left(\frac{1-2p+\pi}{(1-p)^2}\right)+\log\left(\frac{\pi}{p^2}\right)\right]+(p-\pi)\log((1-\rho)^2)\\
&= (1-2p)\log\left(\frac{1-2p+\pi}{(1-p)^2}\right)+(p-\pi)\log((1-\rho)^2)+\pi\log\left(\frac{\pi-2\pi p+\pi^2}{p^2(1-p)^2}\right)
\end{aligned}$$

Using the third equality we can replace $\frac{\pi^2-2p\pi}{p^2(1-p)^2}$ by $(1-\rho)^2-\frac{1}{(1-p)^2}$ and we obtain

$$\begin{aligned}
I(X,Y) &= (1-2p)\log\left(\frac{1-2p+\pi}{(1-p)^2}\right)+(p-\pi)\log((1-\rho)^2)+\\
&\qquad \pi\log\left(\frac{\pi}{p^2(1-p)^2}+(1-\rho)^2-\frac{1}{(1-p)^2}\right)\\
&= (1-2p)\log\left(\frac{1-2p+\pi}{(1-p)^2}\right)+(p-\pi)\log((1-\rho)^2)+\pi\log\left(\frac{\pi-p^2}{p^2(1-p)^2}+(1-\rho)^2\right)\\
&= (1-2p)\log\left(\frac{1-2p+\pi}{(1-p)^2}\right)+(p-\pi)\log((1-\rho)^2)+\pi\log\left(\frac{\rho}{p(1-p)}+(1-\rho)^2\right)\\
&= (1-2p)\log\left(\frac{1-2p+\pi}{(1-p)^2}\right)+p\log((1-\rho)^2)+\\
&\qquad \pi\left[-\log((1-\rho)^2)+\log\left(\frac{\rho}{p(1-p)}+(1-\rho)^2\right)\right]\\
&= (1-2p)\log\left(\frac{1-2p+\pi}{(1-p)^2}\right)+p\log((1-\rho)^2)+\pi\left[\log\left(\frac{(1-\rho)^2+\frac{\rho}{p(1-p)}}{(1-\rho)^2}\right)\right]\\
&= (1-2p)\log\left(\frac{1-2p+\pi}{(1-p)^2}\right)+p\log((1-\rho)^2)+\pi\left[\log\left(1+\frac{\rho}{(1-\rho)^2p(1-p)}\right)\right]\\
&= \log\left(\frac{1-2p+\pi}{(1-p)^2}\right)+p\left[\log((1-\rho)^2)-\log\left(\frac{(1-2p+\pi)^2}{(1-p)^4}\right)\right]+\\
&\qquad \pi\left[\log\left(1+\frac{\rho}{(1-\rho)^2p(1-p)}\right)\right]\\
&= \log\left(\frac{1-2p+\pi}{(1-p)^2}\right)+p\left[\log\left((1-\rho)^2\frac{(1-p)^4}{(1-2p+\pi)^2}\right)\right]+\\
&\qquad \pi\left[\log\left(1+\frac{\rho}{(1-\rho)^2p(1-p)}\right)\right]
\end{aligned}$$

If we denote $\frac{1-2p+\pi}{(1-p)^2} = t$, the mutual information takes the form

$$\begin{aligned} I(X,Y) &= \log t + p\log\left[(1-\rho)^2\frac{1}{t^2}\right] + \pi\log\left[1+\frac{\rho}{(1-\rho)^2p(1-p)}\right] \\ &= \log t + p\log\left[\frac{(1-\rho)^2}{t^2}\right] + \pi\log\left[1+\frac{\rho}{(1-\rho)^2p(1-p)}\right] \\ &= \log t + 2p\log\left[\frac{(1-\rho)}{t}\right] + \pi\log\left[1+\frac{\rho}{(1-\rho)^2p(1-p)}\right] \end{aligned}$$

We can note that, in this case, the mutual information depends on different parameters: t, π, ρ and p. The interpretation of the mutual information is challenging, because the behaviour is determined by the values of the parameters.

[5] proposed a formula that links the mutual information and the covariance for binary sequences. In particular, considering two binary exchangeable random variables X, Y and denoting with $\mathrm{Cov}(X,Y)$ the covariance of X and Y and with $f_X(i) = \Pr(X = i), f_Y(j) = \Pr(Y = j)$, it can be demonstrated that, when

$$\frac{\mathrm{Cov}(X,Y)}{f_X(i)f_Y(j)} \to 0, \text{with } i \in \{0,1\}, j \in \{0,1\},$$

it follows [5]

$$I(X,Y) \approx \frac{1}{2}\left(\frac{\mathrm{Cov}(X,Y)}{f_X(0)f_X(1)}\right)^2 \tag{2.10}$$

Let observe that $\frac{\mathrm{Cov}(X,Y)}{f_X(0)f_X(1)}$ is exactly the correlation coefficient $\rho(X,Y)$, in Equation (2.9), since

$$f_X(0)f_X(1) = \sigma_X^2$$

In summary,

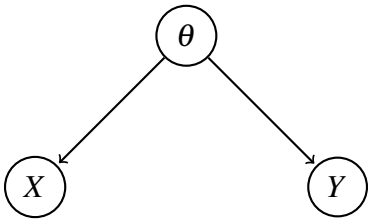
$$I(X,Y) \approx \frac{1}{2}\rho(X,Y)^2, \text{ when } \frac{\mathrm{Cov}(X,Y)}{f_X(i)f_Y(j)} \to 0 \tag{2.11}$$

Two interesting observations from this equation, in this particular case, are that mutual information functions decay to zero at a faster rate than the corresponding correlation functions and that $I(X,Y) = 0 \Leftrightarrow \rho(X,Y) = 0$.

Let analyse, in the following section, the behaviour of mutual information in a specific case of bivariate binary random vector. The same example is proposed, but with the use of the Beta function for the probability distribution.

2.3.1 A specific case

Consider the following Directed Acyclic Graph



where

$$\theta \sim Beta(a,b), \qquad X_{|\theta=p} \sim Bernoulli(p), \qquad Y_{|\theta=p} \sim Bernoulli(p)$$

and let density of θ be

$$f_\theta(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad x \in [0,1], \quad a,b \in (0,\infty)$$

and the conditional density of X be

$$f_{X|\theta=p}(i) = \mathbb{P}(X=i|\theta=p) = p^i(1-p)^{1-i}, \quad i \in \{0,1\}, \quad p \in [0,1]$$

The joint density takes the form

$$\begin{aligned} f_{X,Y,\theta}(i,j,p) &= f_{X|\theta=p}(i) f_{Y|\theta=p}(j) f_\theta(p) \\ &= p^i(1-p)^{1-i} p^j (1-p)^{1-j} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} \end{aligned} \tag{2.12}$$

with $i \in \{0,1\}, j \in \{0,1\}$.
The joint density of X and Y can be obtained by marginalizing out θ.

$$\begin{aligned} f_{X,Y}(i,j) &= \int_0^1 f_{X,Y,\theta}(i,j,p)\mathrm{d}p \\ &= \int_0^1 p^{i+j+a-1}(1-p)^{2-i-j+b-1} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mathrm{d}p \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 p^{i+j+a-1}(1-p)^{1-i-j+b} \mathrm{d}p \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} B(i+j+a, b-i-j+2) \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+i+j)\Gamma(b+2-i-j)}{\Gamma(a+b+2)} \end{aligned} \tag{2.13}$$

It follows that

$$f_{X,Y}(0,0) = \frac{b(b+1)}{(a+b)(a+b+1)}$$
$$f_{X,Y}(0,1) = \frac{ab}{(a+b)(a+b+1)} = f_{X,Y}(1,0)$$
$$f_{X,Y}(1,1) = \frac{a(a+1)}{(a+b)(a+b+1)}$$

Note that the marginal densities of X and Y are identical:

$$f_X(0) = \mathbb{P}(X=0) = \sum_{j\in\{0,1\}} f_{X,Y}(0,j) = \frac{b}{a+b} = f_Y(0) = \mathbb{P}(Y=0) \tag{2.14}$$

$$f_X(1) = \mathbb{P}(X=1) = 1-\mathbb{P}(X=0) = \frac{a}{a+b} = f_Y(1) = \mathbb{P}(Y=1) \tag{2.15}$$

The mutual information can be computed as

$$\begin{aligned}
I(X,Y) &= \sum_{x\in\chi}\sum_{y\in\gamma} f_{X,Y}(x,y)\log\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)} \\
&= f_{X,Y}(0,0)\log\frac{f_{X,Y}(0,0)}{f_X(0)f_Y(0)} + 2f_{X,Y}(0,1)\log\frac{f_{X,Y}(0,1)}{f_X(0)f_Y(1)} + f_{X,Y}(1,1)\log\frac{f_{X,Y}(1,1)}{f_X(1)f_Y(1)} \\
&= \frac{b(b+1)}{(a+b)(a+b+1)}\log\left(\frac{b(b+1)}{(a+b)(a+b+1)}\frac{(a+b)^2}{b^2}\right) + \\
&2\frac{ab}{(a+b)(a+b+1)}\log\left(\frac{ab}{(a+b)(a+b+1)}\frac{(a+b)^2}{ba}\right) + \\
&\frac{a(a+1)}{(a+b)(a+b+1)}\log\left(\frac{a(a+1)}{(a+b)(a+b+1)}\frac{(a+b)^2}{a^2}\right)
\end{aligned}$$

which reduces to

$$I(X,Y) = \log\left(\frac{a+b}{a+b+1}\right) + \frac{b^2+b}{(a+b)(a+b+1)}\log\left(1+\frac{1}{b}\right) + \frac{a^2+a}{(a+b)(a+b+1)}\log\left(1+\frac{1}{a}\right)$$

Some considerations:

- when the density of θ is uniform in the interval $[0,1]$ ($a=b=1$)

$$\begin{aligned}
I(X,Y) &= \log\left(\frac{2}{3}\right) + \frac{2}{6}\log(2) + \frac{2}{6}\log(2) \\
&= \left(1+\frac{4}{6}\right)\log(2) - \log(3) \\
&= \frac{5}{3}\log(2) - \log(3) \approx 0.057
\end{aligned}$$

  It means that the X and Y share a little amount of information.

- Let try to rewrite $I(X,Y)$ in a more interpretable way. Let denote with

$$\begin{cases} p = f_X(1) = \dfrac{a}{a+b} \\ u = a+b \end{cases}$$

It follows that

$$\begin{cases} a = pu \\ b = u(1-p) \end{cases}$$

The mutual information takes the form

$$I_{p,u}(X,Y) = \log\left(\frac{u}{u+1}\right) + \frac{1}{u+1}\left[\left((1-p)^2u + (1-p)\right)\log\left(1+\frac{1}{u(1-p)}\right) + (p^2u+p)\log\left(1+\frac{1}{up}\right)\right]$$

We use R [6] to illustrate the trend of the mutual information as a function of u and p.

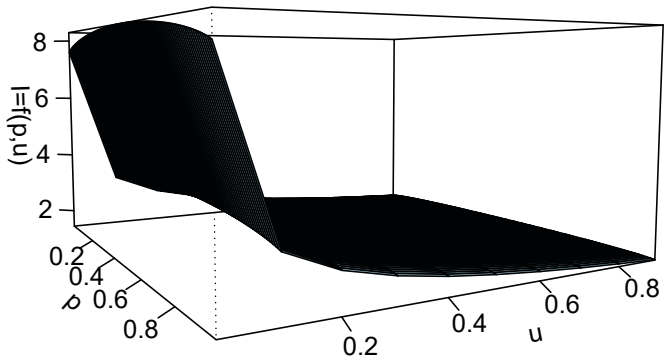


Figure 2.2. The graph shows the trend of the mutual information as a function of u and p, with $0 < u \leq 1$.

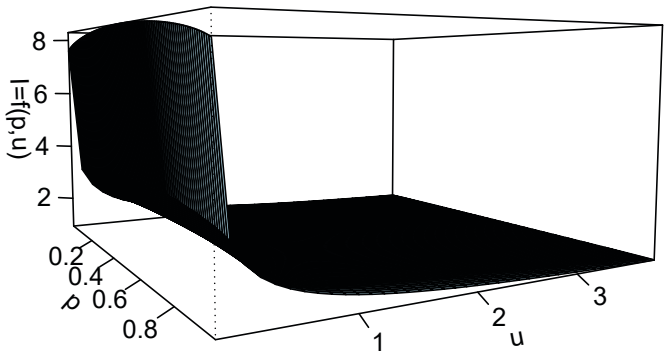


Figure 2.3. The graph shows the trend of the mutual information as a function of u and p, with $0 < u \leq 5$.

From these figures we can see that the mutual information decreases quickly as the sum of the Beta parameters increases. In particular, we have previously shown that when $a = b = 1$, the mutual information takes value close to zero.

Also in this case, as it is the same proposed in the section 2.3, we can approximate (see Equation 2.10) the mutual infomation as

$$\begin{aligned} I(X,Y) &\approx \frac{1}{2}\left(\frac{\mathrm{Cov}(X,Y)}{f_X(0)f_X(1)}\right)^2 \\ &= \frac{1}{2}\left(\frac{f_{(X,Y)}(1,1) - f_X(1)^2}{f_X(0)f_X(1)}\right)^2 \\ &= \frac{1}{2}\left(\frac{1}{a+b+1}\right)^2 \\ &= \frac{1}{2}\left(\frac{1}{u+1}\right)^2 \end{aligned}$$

under the sufficient conditions listed in [5]

$$\frac{\mathrm{Cov}(X,Y)}{f_X(i)f_Y(j)} \to 0, \text{with } i \in \{0,1\}, j \in \{0,1\}$$

that are

$$\begin{cases} \dfrac{\mathrm{Cov}(X,Y)}{f_X(0)f_Y(0)} = \dfrac{f_{(X,Y)}(1,1)-f_X(1)^2}{f_X(0)^2} = \dfrac{a}{b(a+b+1)} \to 0 \\ \dfrac{\mathrm{Cov}(X,Y)}{f_X(1)f_Y(1)} = \dfrac{f_{(X,Y)}(1,1)-f_X(1)^2}{f_X(1)^2} = \dfrac{b}{a(a+b+1)} \to 0 \\ \dfrac{\mathrm{Cov}(X,Y)}{f_X(0)f_Y(1)} = \dfrac{1}{a+b+1} \to 0 \end{cases}$$

With this approximation we can note that the mutual information depends only on the inverse squared sum of the Beta parameters.

Finally, it is of interest to consider a bivariate gaussian mutual information, as it is directly related to the Pearson correlation coefficient. The following section focuses on the Gaussian case.

2.4 The Gaussian case

The mutual information between two Gaussian random variables can be determinated analytically using Equation (2.8).
Let $Z = (X,Y) \sim N(\mu,\Sigma)$ with

$$f(\underline{z}) = \frac{1}{(\sqrt{2\pi})^2|\Sigma|^{\frac{1}{2}}} e^{\frac{-(\underline{z}-\mu)^T\Sigma^{-1}(\underline{z}-\mu)}{2}} \text{ and } \Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}, \ \mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}$$

The joint entropy takes the form

$$\begin{aligned} h(X,Y) &= -\int f(\underline{z})[\frac{-(\underline{z}-\mu)^T\Sigma^{-1}(\underline{z}-\mu)}{2} - \log((\sqrt{2\pi})^2|\Sigma|^{\frac{1}{2}})]\mathrm{d}\underline{z} \\ &= \frac{\mathbb{E}[\sum_{i,j}(z_i-\mu_i)(\Sigma^{-1})_{ij}(z_j-\mu_j)]}{2} + \frac{\log((2\pi)^2|\Sigma|)}{2} \\ &= \frac{\sum_{i,j}\mathbb{E}[(z_i-\mu_i)(z_j-\mu_j)](\Sigma^{-1})_{ij}}{2} + \frac{\log((2\pi)^2|\Sigma|)}{2} \\ &= \frac{\sum_j\sum_i\Sigma_{ji}(\Sigma^{-1})_{ij}}{2} + \frac{\log((2\pi)^2|\Sigma|)}{2} \\ &= \frac{\sum_j(\Sigma\Sigma^{-1})_{jj}}{2} + \frac{\log((2\pi)^2|\Sigma|)}{2} \\ &= \frac{\sum_j I_{jj}}{2} + \frac{\log((2\pi)^2|\Sigma|)}{2} \end{aligned}$$

$$= \frac{2}{2} + \frac{\log((2\pi)^2|\Sigma|)}{2}$$
$$= \frac{\log(2\pi e)^2|\Sigma|}{2}$$

Then, the mutual information can be written as

$$I(X,Y) = \frac{\log(2\pi e\sigma_x^2)}{2} + \frac{\log(2\pi e\sigma_y^2)}{2} - \frac{\log(2\pi e)^2|\Sigma|}{2}$$
$$= -\frac{\log(1-\rho^2)}{2}$$

In this case, is a function of ρ, the Pearson correlation coefficient. In particular, when two Gaussian variables are strictly correlated, $\rho = \pm 1$, $I(X,Y) = \infty$. In contrast, if the two random variables are uncorrelated, $\rho = 0$, then $I(X,Y) = 0$. Indeed, the mutual information is a strictly increasing function of ρ^2, as displayed in Figure 2.4. It follows that in the Gaussian case, the mutual information does not add any information to the linear correlation coefficient ρ.

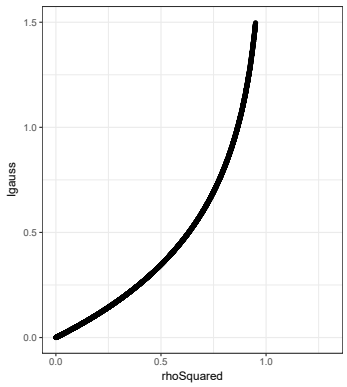


Figure 2.4. The dependence of the mutual information of a bivariate Gaussian vector for different values of ρ^2.

In general the mutual information can detect all types of dependencies, both linear and nonlinear. [5] demonstrates that the mutual information function is capable of capturing the nonlinear dependencies that the covariance might have missed. Indeed, [5] shows that $\mathrm{Cov}(X,Y) = 0 \nRightarrow I = 0$ for ternary sequences in general.

In summary, the mutual information is a more general statistical measure of correlation rather than the Pearson correlation coefficient. It is a dependence measure well defined both for continuous and for categorical variables and we have shown that, in some particular cases and under several sufficient conditions, it can be directly connected to the Pearson correlation coefficient.

Chapter 3

The estimation of entropies and mutual informations

In applications, the data available is often a random sample. To estimate $I(X,Y)$ one begins from N bivariate measurements $(x_i,y_i),\ i = 1,...,N$ each of which are assumed to be $i.i.d.$ (independent identically distributed) realizations of random variables. For two categorical random variables, estimating the joint probability is straightforward, as it consists of counting the number of samples in each combination of categories of the two variables. If two continuous random variables are considered, it becomes more challenging to estimate their joint distribution.

While the problem of the mutual information estimation was extensively studied, it still attracts a lot of attention in the literature. There are two basic approaches to estimation: non-parametric and parametric. Non-parametric estimators are flexible, because they do not assume that the variable is from a known family of distribution, but in contrast they are less powerful (in terms of efficiency and accuracy) than the parametric ones [7]. So the challenge is to find an estimation method that covers both parametric and non-parametric density methodologies and still can be applied to the most if not all applications effectively [7]. Examples of non-parametric entropy estimators are the Kernel Density Estimator [8, 9] and the Kozachenko-Leonenko estimator [10], extended later by [11].

In `R` [6] there are two packages: `infotheo` [12] and `IndepTest` [13] that implement estimators for categorical variables and the Kozachenko-Leonenko estimator, respectively.

The first part of this chapter focuses on the estimation of the mutual information, while the second part considers normalization of mutual information.

3.1 Mutual information of two categorical variables

Let A and B be two categorical random variables. Then, the mutual information can be computed estimating the joint probability from the frequency of observed samples in each combination of variable categories. The estimated mutual information takes the form

$$\begin{aligned}\hat{I}(A,B) &= \sum_{a\in\text{supp}(A)}\sum_{b\in\text{supp}(B)} \hat{p}_{A,B}(a,b)\log\frac{\hat{p}_{A,B}(a,b)}{\hat{p}_A(a)\hat{p}_B(b)} \\ &= \sum_{a\in\text{supp}(A)}\sum_{b\in\text{supp}(B)} \frac{n_{a,b}}{N}\log\frac{\frac{n_{a,b}}{N}}{\frac{n_a}{N}\frac{n_b}{N}} \\ &= \sum_{a\in\text{supp}(A)}\sum_{b\in\text{supp}(B)} \frac{n_{a,b}}{N}\log\frac{Nn_{a,b}}{n_a n_b}\end{aligned}$$

where $n_{a,b}$ is the number of samples with categories a and b, N is the total number of samples, n_a is the number of samples with category a and n_b is the number of samples with category b.

However, the conventional calculation of mutual information based on frequencies of all possible combinations might be not efficient for variables with many categories [14]. In this article, to overcome the inefficiency problem, a recursive partitioning algorithm is proposed. Nevertheless, this algorithm was not considered during the internship. We will not focus on categorical data with a large amount of categories.

3.2 Mutual information of two continuous variables

Let X and Y be two continuous random variables. Below we consider two estimation methods of the mutual information:

1. an estimation using discretization, where the support sets of the random variables are discretized;
2. a non-parametric estimation without discretization.

3.2.1 Estimation using discretization

To estimate the mutual information, we start from the estimation of the entropy and, at a later stage, the estimation of the mutual information can be computed, using Equation (2.4).

Let X be a continuous random variable. We divide the interval of support set into k sub-intervals, called bins and adapt the following notation:

- n_k = the number of samples in bin k;
- N = the total number of samples;
- c = the total number of bins.

The partition of the support set into sub-intervals can be done in `R` with the package `infotheo` [12] using the function *discretize* and one of three different methods of discretization:

- `equalfreq`: division of the interval $[\alpha, \beta]$ into sub-intervals, each having the same number of data points;
- `equalwidth`: division of the interval $[\alpha, \beta]$ into sub-intervals of equal size;
- `globalequalwidth`: uses the same interval width for both random variables.

After the discretization of the variable, the entropy can be estimated using, for example, one of the following methods [15]:

Empirical estimator

$$\hat{H}_{emp} = -\sum_{k=1}^{c} \frac{n_k}{N} \log(\frac{n_k}{N})$$

The empirical estimator is biased [16] and it underestimates the entropy. To adjust it, the Miller-Madow estimator was proposed [15].

Miller-Madow estimator

$$\hat{H}_{mm} = \hat{H}_{emp} + \frac{c-1}{2N}$$

Shrinkage estimator

$$\hat{H}_{sk} = -\sum_{k=1}^{c} \hat{p_\lambda}(n_k) \log(\hat{p_\lambda}(n_k))$$

where

$$\hat{p_\lambda}(n_k) = \lambda \frac{1}{c} + (1-\lambda)\frac{n_k}{N}$$

and the weighting parameter λ is estimated by minimizing:

$$\lambda^* = \arg\min_{\lambda \in [0,1]} \mathbb{E}[(\sum_{k \in c} \hat{p_\lambda}(n_k) - p(n_k))^2]$$

Note that if the parameter $\lambda \to 0$ the shrinkage estimator converges to the empirical one. Instead, if $\lambda \to 1$ the probability $\hat{p_\lambda}(n_k)$ follows a uniform distribution.

Schurmann-Grassberger estimator

The Schurmann-Grassberger estimator uses the Dirichlet probability distribution as a conjugate prior for the likelihood given by the empirical estimator. The prior parameter is chosen as $\frac{1}{c}$. Then, the entropy is estimated as follow:

$$\hat{H}_{dir} = -\sum_{k=1}^{c} \hat{p}_k \log \hat{p}_k$$

where

$$\hat{p}_k = \frac{n_k + \frac{1}{c}}{N+1}$$

3.2.2 Estimation without discretization

Kernel Density Estimator

The kernel density estimation was introduced by [8, 9]. The general form of a Kernel Density Estimator (KDE) in d dimensions is

$$\hat{p}(\mathrm{x}) = \frac{1}{Nh^d} \sum_{k=1}^{N} K\left(\frac{\mathrm{x} - \mathrm{x}_k}{h}\right)$$

where $K(\mathrm{x})$ is the kernel function, which is required to integrate to one [9], x is a d-dimensional random vector, N is the number of samples and h is the kernel width. The performance of KDE estimators strongly depends on the choice of the kernel width.

The mutual information, using the KDE estimator, can be estimated as

$$\hat{I}(X,Y) = \sum_{x\in\chi}\sum_{y\in\gamma} \hat{p}_{X,Y}(x,y) \log \frac{\hat{p}_{X,Y}(x,y)}{\hat{p}_X(x)\hat{p}_Y(y)} \tag{3.1}$$

Several articles, in the literature, focused on KDE estimator and the choice of the optimal bandwidth. However, [17] states that the Kozachenko-Leonenko estimator is computationally more effective and stable than the KDE estimator. During the internship, we have decided to focus on the Kozachenko-Leonenko estimator.

Kozachenko-Leonenko estimator

A non-parametric entropy estimator is the Kozachenko-Leonenko [10]. This was subsequently modified by [18] to estimate the mutual information.

In particular, in the package `IndepTest` [13], the function *mutinfo* computes the mutual information using the estimator $I^{(1)}(X,Y)$ described by Kraskov [18].

Consider two continuous random variables X and Y and the space $Z = (X,Y)$. The point in the space is $z_i = (x_i, y_i)$. The norm in the metric space is defined as

$$\|z_i - z_j\| = \max\{\|x_i - x_j\|, \|y_i - y_j\|\}$$

For each point z_i we can compute the distances $d_{i,j} = \|z_i - z_j\|, \forall j \neq i$ and rank the neighbours of z_i by distance: $d_{i,j_1} \leq d_{i,j_2} \leq d_{i,j_3} \leq ...$
Following the original work, we use the notation below:

- $\frac{\varepsilon(i)}{2}$ is the distance between z_i and his k-nearest neighbour.
- $\frac{\varepsilon_x(i)}{2}$ and $\frac{\varepsilon_y(i)}{2}$ are the distance as above projected into the subspace X and Y, it follows that $\varepsilon(i) = \max\{\varepsilon_x(i), \varepsilon_y(i)\}$
- $n_x(i)$ is the number of points x_j whose distance from x_i is strictly less than $\frac{\varepsilon(i)}{2}$.

An example of the space Z is given in Figure 3.1.

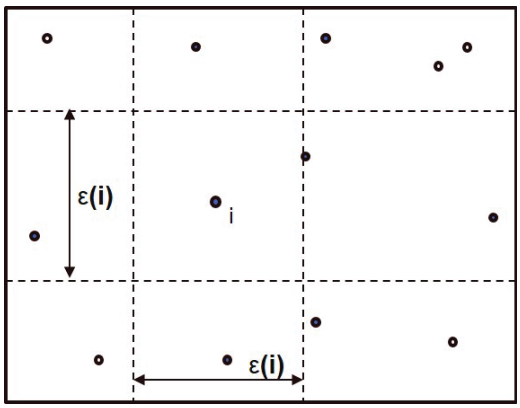


Figure 3.1. k=1, $n_x(i) = 2$ and $n_y(i) = 2$.

The Kozachenko-Leonenko estimator is defined as [18]

$$\hat{H}(X) = -\psi(k) + \psi(N) + \log c_d + \frac{d}{N}\sum_{i=1}^{N} \log \varepsilon(i) \tag{3.2}$$

where ψ is the digamma function, defined as $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\ln(\Gamma(x)) = \frac{\Gamma'(x)}{\Gamma(x)}$, d is the dimension of X, c_d is the volume of the d-dimensional unit ball and N is the total number of samples.

To avoid the different bias obtained in $\hat{H}(X), \hat{H}(Y)$ and $\hat{H}(X,Y)$, [18] proposed to estimate the entropy in the following way

$$\hat{H}(X) = -\frac{1}{N}\sum_{i=1}^{N}\psi[n_x(i)+1] + \psi(N) + \log c_{d_x} + \frac{d_x}{N}\sum_{i=1}^{N}\log\varepsilon(i) \tag{3.3}$$

and the Kraskov estimator for the mutual information takes the form

$$\hat{I}(X,Y) = \psi(k) + \psi(N) - \frac{1}{N}\sum_{i=1}^{N}\psi[n_x(i)+1] + \psi[n_y(i)+1] \tag{3.4}$$

Recent papers have explored the main properties of the Kraskov estimator (KSG estimator). Specifically, [19] demonstrates the consistency of the estimator and proposed a bias-improved KSG estimator.

3.3 Mutual information of a categorical and a continuous variable

Let X be a continuous random variable and A be a categorical random variable defined by a number of distinct classes. The estimator of the mutual information of categorical and continuous random variables is proposed in [20].

Let

- X be a continuous random variable and A be a categorical random variable with L classes
- $\Pr(A = a_l) = \dfrac{n_l}{N}$
- $\dfrac{\varepsilon(n,k)}{2}$ be the distance between x_n and his k-nearest neighbour.
- $\dfrac{\varepsilon_l(n,k)}{2}$ be the distance between x_n and his k-nearest neighbour, but the set of neighbours of x_n is computed using the data having class a_l only.

The estimator is

$$\hat{I}(X,A) = \psi(N) - \frac{1}{N}\sum_{l=1}^{L} n_l\psi(n_l) + \frac{d}{N}\left(\sum_{n=1}^{N}\log\varepsilon(n,k) - \sum_{l=1}^{L}\sum_{y\in y_l}\log\varepsilon_l(n,k)\right) \tag{3.5}$$

where d is the dimension of X, ψ is the digamma function and N the total number of samples.

It was implemented in the software `R` by the author, during the internship. For details, see the code in Appendix B.

3.4 The normalization

The mutual information can take values from zero to infinity. To compute hierarchical clustering among the variables, a distance or a measure of dissimilarity defined on the same scale is needed. To define a measure of distance, the mutual information has to be upper bounded.

We start by separating the case of two categorical random variables and the case of two continuous random variables.

3.4.1 Categorical random variables

Let A and B be two categorical random variables. [21] defines the following measure of similarity using the disequality e) from the properties listed in Chapter 2:

$$0 \leq I(A,B) = H(A) - H(A|B) \leq \min(H(A), H(B))$$

and takes the normalized mutual information as:

$$I(A,B)_{norm} = \frac{I(A,B)}{\min(H(A), H(B))}, \qquad 0 \leq I(A,B)_{norm} \leq 1 \tag{3.6}$$

Then, the dissimilarity measure takes the form

$$d'(A,B) = 1 - \frac{I(A,B)}{\min(H(A), H(B))} \tag{3.7}$$

However, it can be demonstrated that $d'(A,B)$ is not a distance, as it does not fulfill the identity of indiscernibles: the distance can reach zero even if the entropies $H(A)$ and $H(B)$ are not identical. For example, if we consider

$$H(A) = \min(H(A), H(B))$$

then

$$d'(A,B) = 1 - \frac{I(A,B)}{H(A)} = 0 \;\Leftrightarrow\; H(A) = I(A,B) = H(A) + H(B) - H(A,B)$$

which implies $H(B) = H(A,B)$ and not $A = B$. However, this measure is useful in *content recognition* [21], where a metric is not necessary.

The second normalized version is:

$$I(A,B)_{norm} = \frac{I(A,B)}{H(A,B)}, \qquad 0 \leq I(A,B)_{norm} \leq 1 \tag{3.8}$$

Indeed, if B coincides with A, then

$$I(A,A) = H(A), \;\; H(A,A) = H(A)$$

and

$$I(A,A)_{norm} = 1$$

The distance measure can be defined as

$$d(A,B) = 1 - \frac{I(A,B)}{H(A,B)} \tag{3.9}$$

and it can be demonstrated that this measure is a metric [22].

3.4.2 Continuous random variables

Considering continuous random variables, the differential entropy can be negative. Therefore, the normalization cannot be well defined.

A normalized version of the mutual information was proposed by [23] and was subsequently used by [24, 25].

Let X and Y be two continuous random variables, then

$$r(X,Y) = \sqrt{1 - \exp^{-2I(X,Y)}} \tag{3.10}$$

is called the information coefficient of correlation. Note that if X and Y are normally distributed, $r(X,Y) = |\rho|$, where ρ is the Pearson correlation coefficient. Indeed:

$$r(X,Y) = \sqrt{1 - \exp^{-2I(X,Y)}} = \sqrt{1 - \exp^{-2(-\frac{ln(1-\rho^2)}{2})}} \quad = \sqrt{1 - (1-\rho^2)} = |\rho|.$$

3.4.3 Continuous and categorical random variables

Dealing with mixed dataset, containing both categorical and continuous variables, there is also the case of the mutual information between a continuous and a categorical random variables.

We do not focus on the normalization of the estimator of mutual information (Eq. 3.5), because, as we will see in Chapter 5, the dataset of interest will be split in two parts: a group with continuous variables only and a group with categorical variables only. In this way, just two distance measures (Eq. 3.9 and the dissimilarity from Eq. 3.10) are needed and they will be analized in the simulation study in Chapter 5.

To summarize, the mutual information has been widely studied in the past and several estimators were proposed during the years. During the internship we have decided to use one of them and we have analysed the results obtained.
Next chapters regard the description of the work done during the internship and the application of mutual information to an artificial example.

Chapter 4

A solution to explore associations between variables

In literature, [21, 22, 25] use the mutual information for hierarchical clustering and classification problems. In particular [22] proposed a hierarchical clustering based on the mutual information and applied it to two datasets containing continuous variables. It uses the mutual information as a proxy to group objects into clusters. Also [25] suggests an agglomerative hierarchical clustering to study interdependencies among continuous variables. In Tetra Pak we have decided to explore the data using a similar method described in this chapter. The second part of the chapter focuses on another approach used in the literature to group mixed variables.

4.1 A description of the method

In the following section the explanation of the work done during the internship is given. In Tetra Pak the described method was applied to real datasets and the results obtained were verified and used for development activities. However, for confidentiality reasons, the real dataset cannot be reported in this document. For this, in Chapter 5, a simulation study is reported.

4.1.1 The idea elaborated during the internship

The idea, elaborated during the internship, is to solve the points described in the section 1.3 in this way:

- Compute pairwise correlations among all pairs of variables using a measure of dependence:
  - considering mixed datasets with categorical and continuous variables, we use a unique estimator of the mutual information. For continuous variables the `equalfreq` discretization is used. The mutual information is estimated using the `Shurmann-Grassberger` estimator.
- Compute hierarchical clustering among variables using the distance (3.9) and the complete-linkage clustering. Denoting with V_1 and V_2 two clusters, the complete-linkage is the distance between V_1 and V_2 and it is defined as

$$D(V_1, V_2) = \max\{d(v_1, v_2) : v_1 \in V_1, v_2 \in V_2\}$$

- Given a variable of interest Y, for each cluster, find the variable that shares the most information with Y:
  - two different cases are considered:
    a) Y is continuous: if the response variable is continuous, the mutual information between Y and the others variables can be estimated using the *Kraskov* estimator [18] and the estimator proposed in Equation (3.5) for the case of a continuous and a categorical variables.
    b) Y is categorical: the mutual information between the response variable and the others is estimated using the `Shurmann-Grassberger` estimator.

As previously mentioned, the proposed method cannot be presented, in the thesis, for the real data stored in Tetra Pak. Therefore, to have a comparison measure on a fictional dataset, another method to explore associations is proposed in the following section.

4.2 An alternative approach to explore mixed dataset

In literature another method is proposed for the clustering of variables and it is implemented in `R` in the package `ClustOfVar` [26].

The method developed in the package works with all types of variables (both continuous and categorical) and there is a function *hclustvar* that computes a hierarchical clustering among the features of the dataset. The method is based on PCAMIX [27], a principal component method for both continuous and categorical variables. The function *hclustvar* is used to compare the results of the two different methods (PCAMIX and the mutual information approach) on an artificial dataset.

4.2.1 A brief description of the ClustOfVar package

The clustering method used in the `ClustOfVar` package aims to maximize the homogenity criterion: a cluster of variables is considered to be homogenous if the variables into the cluster are strongly linked to a central quantitative synthetic variable. The central quantitative variable of a cluster is the first principal component of PCAMIX applied to all the variables in the cluster.

Let us consider two sets:

- $X = \{x_1, ..., x_{p_1}\}$, a set of continuous variables; X is a matrix of dimension $n \times p_1$, where n is the number of the observations;
- $Y = \{y_1, ..., y_{p_2}\}$, a set of categorical variables; Y is a matrix of dimension $n \times p_2$.

For simplicity, denote the j-th column of X by x_j, the j-th column of Y by y_j and the set of categories of y_j by M_j. Let $P_k = (C_1, ..., C_k)$ be a partition into K clusters of the $p = p_1 + p_2$ variables.

The synthetic variable of a cluster C_k is defined below

Definition 4.1.

$$c_k = \underset{u \in \mathbb{R}^n}{\arg\max} \{ \sum_{x_j \in C_k} r^2_{u,x_j} + \sum_{y_j \in C_k} \eta^2_{u|y_j} \} \tag{4.1}$$

where r^2 is the squared Pearson correlation coefficient and η^2 denotes the correlation ratio, that is defined as

$$\eta^2_{u|y_j} = \frac{\sum_{s \in M_j} n_s (\overline{u}_s - \overline{u})^2}{\sum_{i=1}^{n} (u_i - \overline{u})^2} \tag{4.2}$$

where n_s is the frequency of category s, $\overline{u}_s$ is the mean value of u calculated on the observations belonging to category s and $\overline{u}$ is the mean of u.

Let us define the homogeneity H of a cluster C_k.

Definition 4.2.

$$H(C_k) = \sum_{x_j \in C_k} r^2_{x_j,c_k} + \sum_{y_j \in C_k} \eta^2_{c_k|y_j} \tag{4.3}$$

The aim is to find a partition of a set of continuous and categorical variables such that the variables within a cluster are strongly related to each other. In the package a hierarchical clustering algorithm is proposed.

The hierarchical clustering algorithm

1. Step $i = 0$: start with the partition in p clusters.

2. Step $i = 1, ..., p-2$: aggregate two clusters of the partition in $p-i+1$ clusters to get a new partition in $p-i$ clusters. The clusters A and B are chosen with the smallest dissimilarity d defined as

$$d(C_1, C_2) = H(C_1) + H(C_2) - H(C_1 \cup C_2) \tag{4.4}$$

3. Step $i = p-1$: stop. The partition in one cluster is obtained.

In summary, in this chapter the idea elaborated during the internship was presented. However, for privacy conditions, the tests and the results analysed during the internship cannot be stated in the document. For this reason, the Chapter 5 presents a simulation study and the application of the previous approaches to an artifical dataset created by the author.

Chapter 5

Simulation study

The aim of this chapter is to simulate an artificial dataset in order to investigate the behaviour of exploratory method. Initially, we analyse the first two points listed in 1.3:

1. Compute pairwise correlations among all variables using a measure of dependence;
2. Compute hierarchical clustering among variables.

The last point, that is to find, for each cluster, the variable that shares the most information with a variable of interest Y, is taken into account in the second part of this chapter.

5.1 Artificial dataset

To have a prior idea of the clusters and the dependencies among variables, an artificial dataset was simulated. The code used is reported in Appendix A. The artificial dataset contains both continuous and categorical variables. In particular, there are both linear and nonlinear functional dependencies among the continuous variables and some of the categorical variables are linked to the continuous ones. For example, the categorical variable *exponentialSign* assumes value "+" if the continuous random variable *exponential* is greater than three and the value "−" if the continuous random variable *exponential* is less or equal to three.
The response variable is *risp*.

The exploration analysis is organized as follow:

- the dataset is divided into two subsets: one containing the continuous variables only and the other the categorical ones;
- the full dataset is considered.

5.1.1 Continuous variables

Let us take into account only the continuous variables and let investigate the associations among them.

The pairwise distributions among the continuous variables are shown in the following figure:

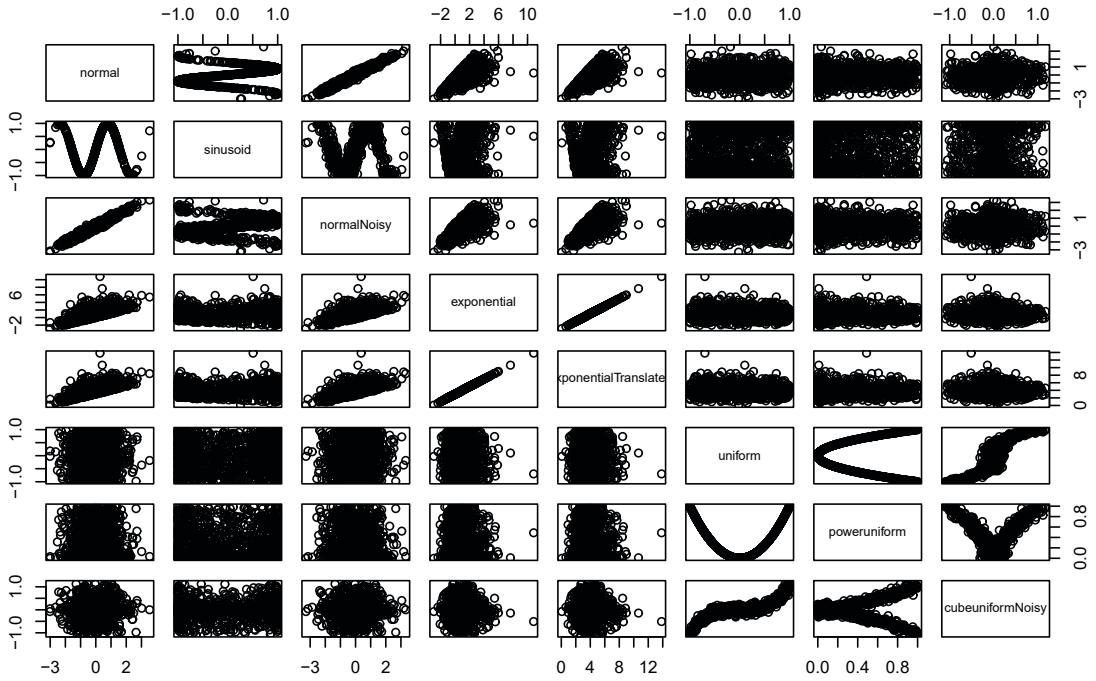


Figure 5.1. Distributions of the continuous variables.

Figure 5.1 shows that there are three main clusters among the variables. Specifically, we expect to find the following groups of variables:

- *normal, sinusoid, normalNoisy*;
- *exponential, exponentialTranslated*;
- *uniform, poweruniform, cubeuniformNoisy*.

To search these main dependencies, we use:

- the Kraskov estimator, Eq. (3.4), to compute correlations pairwise.

- $dist(X,Y) = 1 - r(X,Y) = 1 - \sqrt{1 - \exp^{-2I(X,Y)}}$, as the distance between two random variables, where $r(X,Y)$ is the information coefficient of correlation given in Equation (3.10).

The hierarchical tree obtained is given in Figure 5.2

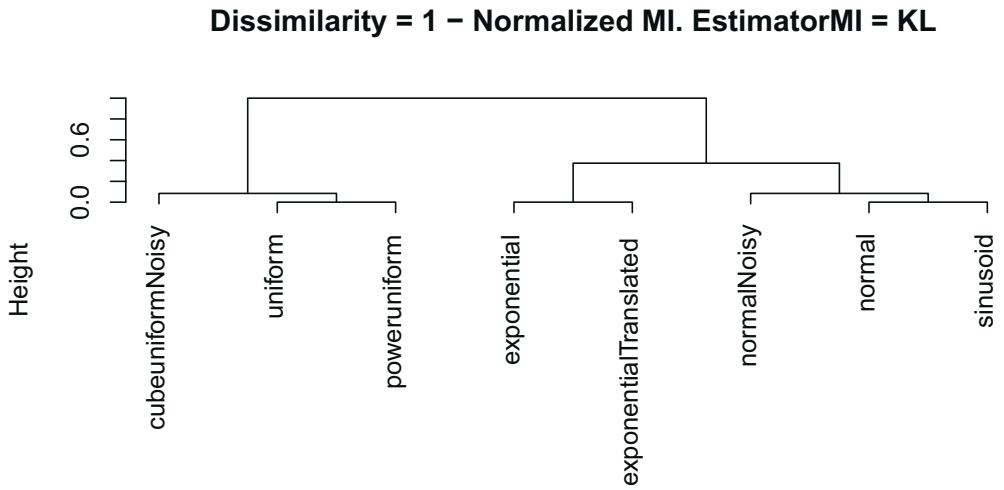


Figure 5.2. Clustering of continuous variables.

Using the proposed approach, all the clusters are consistent with the definitions of the variables and the groups found coincide with the expected ones.

5.1.2 Categorical variables

Let now take into account only the categorical variables of the dataset. To create the hierarchical clustering we use the method described in the Chapter 4.

Looking at the definitions of the variables in Appendix A, we expect to find the following groups of variables:

- *capitalLetter, lowercaseLetter, letterGroup*;
- *exponentialSign, exponentialSignLetter*;
- *letterUniform, uniformdiscretize*;
- *sinusoidDomainSign*.

The hierarchical tree is presented in Figure 5.3

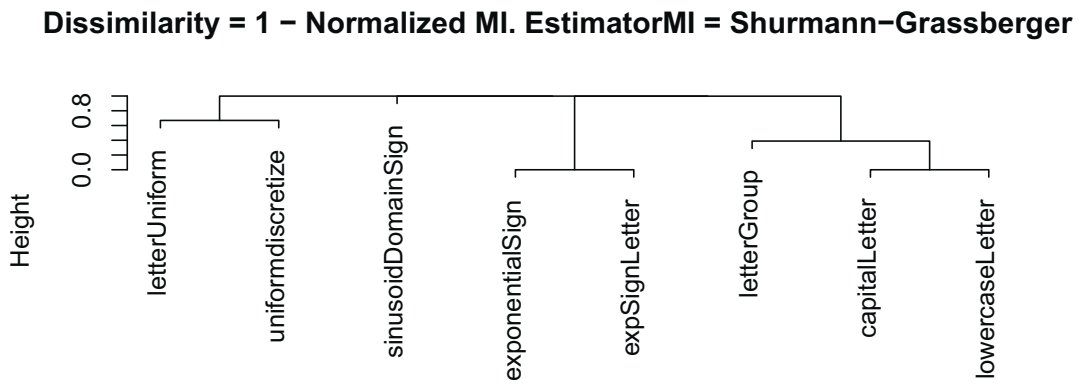


Figure 5.3. Clustering of categorical variables.

Clearly, all the clusters found coincide with the expected ones.

5.1.3 Full Dataset

Let us consider the full dataset and also in this case, the method described in Chapter 4 is used. Let us compute pairwise correlations using the two different methods of discretization which are described in the section 3.2.1: `equalfreq` and `equalwidth`.

Looking at the variables in Appendix A, we expect that the categorical variables defined from the continuous ones fall into the same cluster. In particular:

- *sinusoid, sinusoidDomainSign*;
- *exponential, exponentialSign, exponentialSignLetter*;
- *uniform, letterUniform, uniformdiscretize*.

The hierarchical tree obtained using the discretization `equalfreq` is given in Figure 5.4

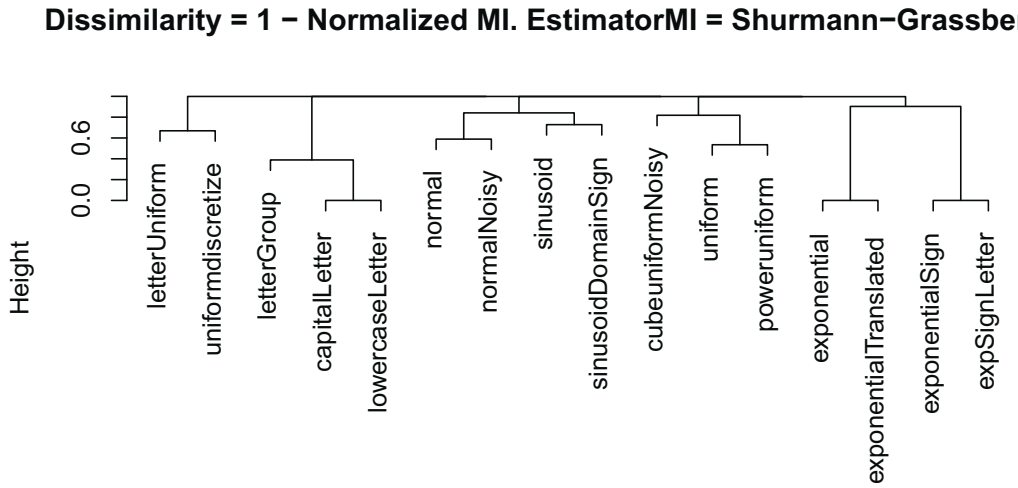


Figure 5.4. Clustering of variables of the artificial dataset.

The hierarchical tree obtained using the discretization `equalwidth` is provided in Figure 5.5

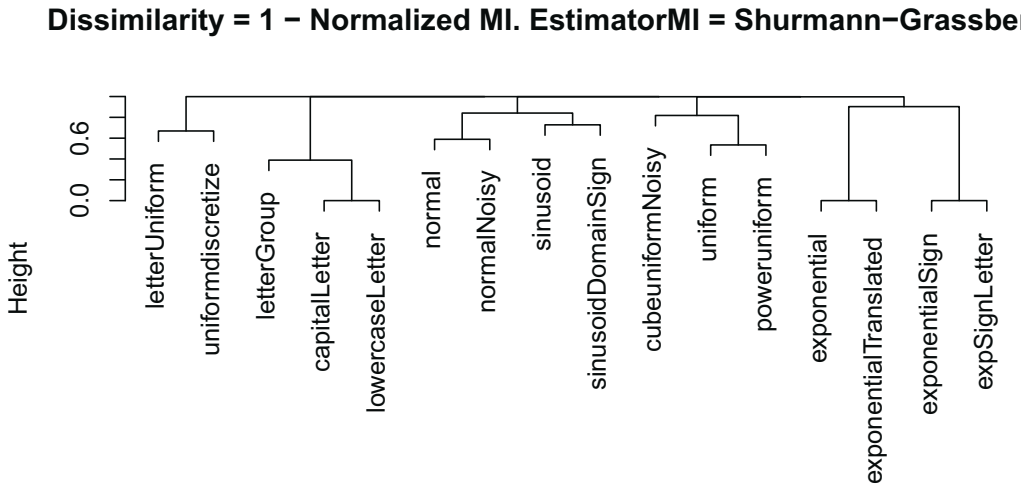


Figure 5.5. Clustering of variables of the artificial dataset.

We compare these hierarchical trees with the hierarchical tree created by the function

hclustvar in the package `ClusOfVar`.

The hierarchical tree obtained is demonstrated in Figure 5.6

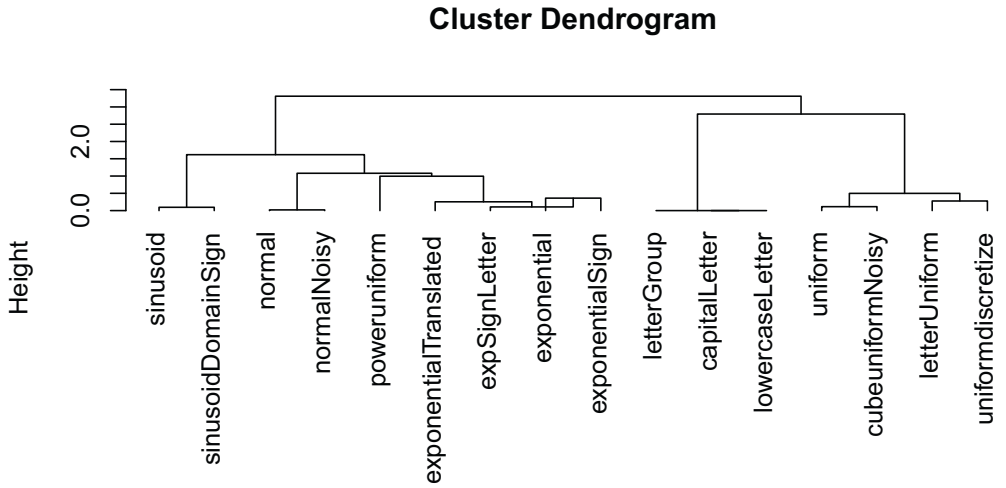


Figure 5.6. Clustering of variables of the artificial dataset using the `ClustOfVar` package.

Some considerations:

- the choice of the discretization method in the first two hierarchical trees does not influence the result. The clusters obtained are exactly the same.
- Having a look at Figure 5.4 and Figure 5.6, it can be noted that the main clusters found are the same. The difference is in the cluster *letterUniform* and *uniformdiscretize*. Using the first method (see Figure 5.4), the cluster is isolated from the others, while using the `ClustOfVar` package (Figure 5.6) the cluster is correctly connected to the cluster of the variable *uniform*.
- In the second method, the non-linear dependencies are difficult to find out, due to the fact that in the `ClustOfVar` package the correlation is computed using the Pearson correlation coefficient. Indeed a strong dependency is detected between the variable *uniform* and the variable *cubeuniformNoisy* (see Figure 5.6). Moreover the variable *poweruniform* is wrongly associated in the cluster of exponential variables. Instead, the first method, is better in this case, because it combines correctly the variable *poweruniform* and it finds out a stronger correlation between *uniform* and *poweruniform* than *uniform* and *cubeuniformNoisy*.

In general, paying attention to the results obtained, the method proposed finds out the main expected correlations. It means that, in this case, the loss of information due

to the discretization of the continuous variables does not influence the conclusion of the analysis. However, this analysis refers only to one realization of the simulated dataset. Therefore, the final results could be affected by chance.

In Appendix C the results of ten simulations are reported. To discover the clusters obtained the author has to look at the hierarchical trees, so the number of simulations is restricted because it becomes difficult to compare hundred of hierarchical trees manually. Observing all the simulations we can say that, in general, the associations found coincide with the expected ones. The hierarchical trees obtained using the mutual information approach are almost the same in all the simulations, while the trees obtained using the `ClustOfVar` package present some differences:

- the variable *poweruniform* is wrongly placed in all the simulations;
- in some simulations the cluster formed by the variables *normal, normalNoisy* is strongly connected to the cluster formed by *exponential, exponentialTranslated* and sometimes is connected with the cluster *exponential*, *exponentialTranslated*, *exponentialSign*, *expSignLetter*.

In conclusion, the method based on mutual information, in comparison with the hierarchical algorithm proposed in the `ClustOfVar` package, works pretty well, because it finds all the expected associations among the variables. However, particular attention must be paid for the noise term. All the conclusions refer to simulations not affected by noise.

5.1.4 Artificial dataset affected by noise term

In the following section we analyse the results on a noisy dataset. In particular, we use the artificial dataset and we include a term of noise. Specifically, we add a uniform random term to the continuous variables of the dataset and we compare each result with the trees obtained with the `ClustOfVar` package. All the results obtained by one realization of the dataset are reported in the Appendix D.

Having a look at the figures in the Appendix D, we can say that the main expected clusters are always found. However there are some differences between the proposed method and the `ClustOfVar` package.

- Looking at the figures obtained with the method based on mutual information: the variable *uniform*, with the increase of noise, is associated to the variable *cubeuniformNoisy* instead of being correlated to the variable *poweruniform*. The other clusters remain the same with the increase of the noise term.
- Looking at the figures obtained with the method based on the `ClustOfVar` package:

the principal clusters are always found; the only problem is the position of the variable *poweruniform*, that is, most of the times, wrongly collocated.

We can conclude that the add of noise does not influence so much the groups found.

In the following section let consider a variable of interest and let investigate how the method based on mutual information works.

5.1.5 Hierarchical clustering with the variable of interest

In this section we analyse the third point cited in the section 1.3:

3. Given a variable of interest Y, for each cluster find the variable that shares the most information with Y

The variable of interest Y, in this simulation, is *risp*. The variable *risp* is a discretization of the variable *normal*. Look at Appendix A for details.
In particular, Y assumes values:

- "sdout" if $normal >= 1$ or $normal <= -1$
- "sdin" if $-1 < normal < 1$

Therefore we expect *normal* to be the variable that shares most information with Y.
The hierarchical tree, obtained using the method proposed in the previous chapter, is reported in Figure 5.7.

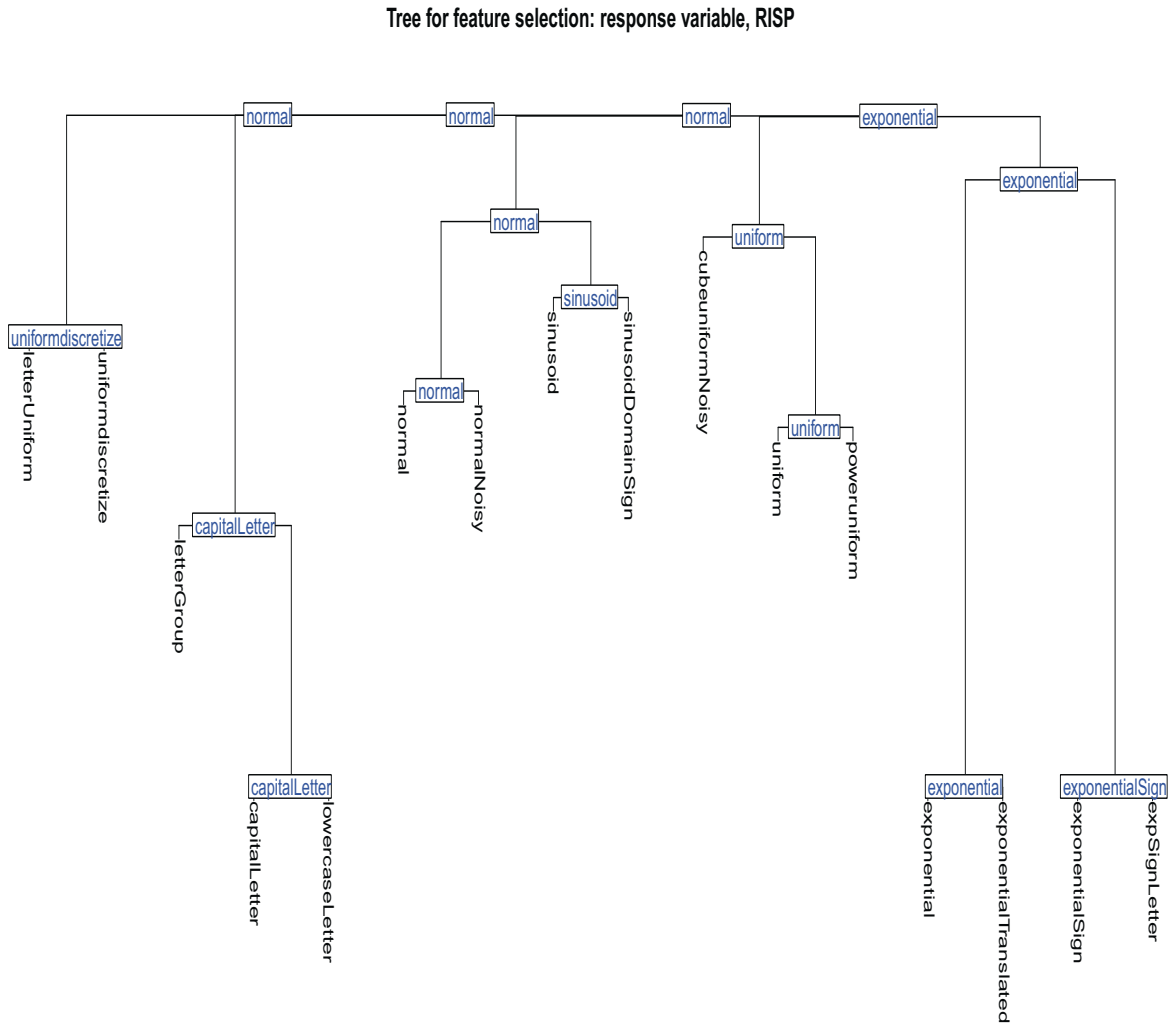


Figure 5.7. Clustering of variables with the response variable.

For each cluster the most representative variable is chosen. The root of the tree represents the variable that shares the most information with the response Y.
The method used finds that the variable *normal* explains lots of the information contained in Y. The result obtained coincides with our expectations.

Chapter 6

Conclusion and future works

The aim of this thesis was to find a method to explore data and discover correlations among the variables of a dataset.
A possible solution to group variables of datasets is proposed, but there are several points that could be explored in the future works, for example:

- more exhaustive testings on real data are necessary;
- several testings with other distance measures;
- study if it is possible to define a measure, based on information theory, that could be used for both categorical and continuous variables;
- understand how to compare results using two or more different normalization measures.

The work presented in this document can be a basis for future works focused on searching a set of uncorrelated variables in order to build predictive models to explain variables of interest. This type of works, in the specific case of Tetra Pak, would be an important step in the direction of modeling and predicting the risk of having cases of unsterility.

Appendix A

Artificial dataset

```
i <- 1000
#numericData
normal <- rnorm(i,0,1)
sinusoid <- sin(2*normal)
normalNoisy <- normal + rnorm(i,0,0.2)
uniform <- runif(i,-1,1)
poweruniform <- uniform^2
cubeuniformNoisy <- uniform^3 + rnorm(i,0,0.1)
exponential <- rexp(i,1) + normal
exponentialTranslated <- exponential + 3
#categoricalData
capitalLetter <- rep(c("A","B", "C","D"), times = c(5,200,695,100), replace =
    TRUE)
lowercaseLetter <- replace(capitalLetter, which(capitalLetter == "A"), "a")
lowercaseLetter <- replace(lowercaseLetter, which(lowercaseLetter == "B"), "b")
lowercaseLetter <- replace(lowercaseLetter, which(lowercaseLetter == "C"), "c")
lowercaseLetter <- replace(lowercaseLetter, which(lowercaseLetter == "D"), "d")
sinusoidDomainSign <- ifelse(sinusoid<0, "NEG", "POS")
letterGroup <- ifelse(capitalLetter == "A" | capitalLetter == "B", "0", "1")
exponentialSign <- ifelse(exponential>3, "+", "-")
expSignLetter <- ifelse(exponentialSign=="+", "MORETHAN3", "LESSTHAN3")
letterUniform <- ifelse(uniform>0 & capitalLetter=="C", "0", "1")
uniformdiscretize <- ifelse(uniform<=0, "UNINEG", "UNIPOS")
risp <- ifelse(normal >=1 | normal <=-1 , "sdout", "sdin")

dataset <- data.frame(normal,sinusoid,normalNoisy,exponential,
    exponentialTranslated, uniform, poweruniform, capitalLetter, lowercaseLetter,
    sinusoidDomainSign, letterGroup,exponentialSign,expSignLetter,
    cubeuniformNoisy,letterUniform, risp,uniformdiscretize)
```

Appendix B

R codes

Mutual information between a categorical and a continuous random variable

```
miDiscCont <- function(x,k)
{
  dim = dim(as.matrix(x))
  n = dim[1]
  d = dim[2]-1

  #categorical feature
  y <- x[,which(sapply(x,is.factor))]
  #continuous feature
  r <- x[,which(sapply(x,is.numeric))]

  Table <- as.data.frame(table(y[,drop = TRUE]))
  # k has to be always smaller than the minimum number of nl
  if (k >= min(table(y[,drop = TRUE]))){

        print(paste("The problematic class label is",
            Table[Table$Freq==min(table(y[,drop = TRUE])),1], sep = " "))
        stop("K has to be less than the smaller class label: change the value of
            K or eliminate the problematic rows")
    }

    logepsl <- array(0,c(1, length(table(y[,drop = TRUE]))))
    nl <- as.vector(table(y[,drop = TRUE]))
    dig <- as.vector(digamma(nl))
    eps <- 2*knn.dist(r, k= k)[,k]
    class <- as.character(Table[,1])
    for (i in 1:length(table(y[,drop = TRUE]))){

      epsl <- 2*knn.dist(x[which(y == class[i]), which(sapply(x,is.numeric))], k=
          k)[,k]
      logepsl[i] <- sum(log(epsl))
    }

    MI <- digamma(n) - (1/n) *sum(nl*dig) + (d/n) * (sum(log(eps)) - sum(logepsl))

  if(MI < 0)
    MI <- 0
  return(MI)
}
```

Appendix C

Multiple simulations

The appendix contains the proportion of correctly identified associations for ten different simulations of the artificial dataset. The correct clusters, which coincide with the definitions of the variables, are reported in italics.

In particular with the mutual information approach we obtained the following clusters in all the simulations.

Cluster	Proportion of clusters
capitalLetter, lowercaseLetter, letterGroup	$\frac{10}{10}$
exponentialSign, exponentialSignLetter	$\frac{10}{10}$
exponential, exponentialTranslated	$\frac{10}{10}$
sinusoid, sinusoidDomainSign	$\frac{10}{10}$
normal, normalNoisy	$\frac{10}{10}$
letterUniform, uniformdiscretize	$\frac{10}{10}$
uniform, poweruniform, cubeuniformNoisy	$\frac{10}{10}$

Table C.1. Proportion of obtained clusters on ten simulations.

Note that the associations letterUniform, uniformdiscretize and uniform, poweruniform, cubeuniformNoisy are correct, but they were expected to be in the same cluster and not separated.

However, using the `ClustOfVar` package we obtain the following clusters.

Cluster	Proportion of clusters
capitalLetter, lowercaseLetter, letterGroup	$\frac{9}{10}$
uniform, cubeuniformNoisy, letterUniform, uniformdiscretize	$\frac{8}{10}$
exponential, exponentialTranslated, exponentialSign, exponentialSignLetter	$\frac{6}{10}$
sinusoid, sinusoidDomainSign	$\frac{6}{10}$
normal, normalNoisy	$\frac{4}{10}$
normal, normalNoisy, exponential, exponentialTranslated	$\frac{4}{10}$
poweruniform, sinusoid, sinusoidDomainSign	$\frac{4}{10}$
exponentialSign, exponentialSignLetter	$\frac{3}{10}$
uniform, poweruniform, cubeuniformNoisy, letterUniform, uniformdiscretize	$\frac{2}{10}$
normal, normalNoisy, poweruniform	$\frac{2}{10}$
exponentialSign, exponentialSignLetter, poweruniform	$\frac{1}{10}$
capitalLetter, lowercaseLetter, letterGroup, poweruniform	$\frac{1}{10}$

Table C.2. Proportion of obtained clusters on ten simulations.

Appendix D

Artificial dataset with noise

In the following pages five different comparisons among the hierarchical trees are reported. In each page the trees refer to a specific dataset obtained using the artificial dataset and adding a uniform term of noise.

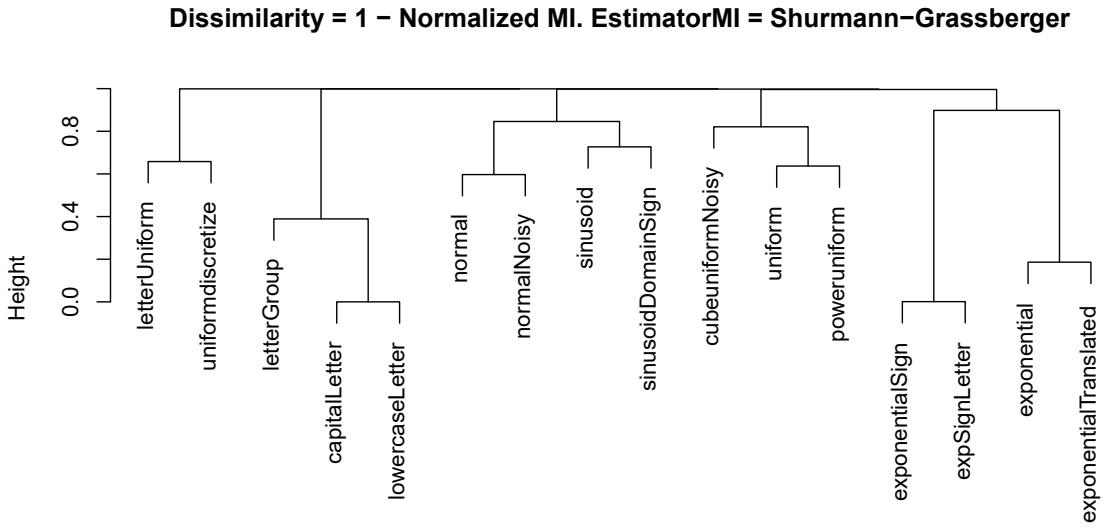


Figure D.1. Clustering of variables for dataset with add of noise $N \sim U(0.01, 0.1)$.

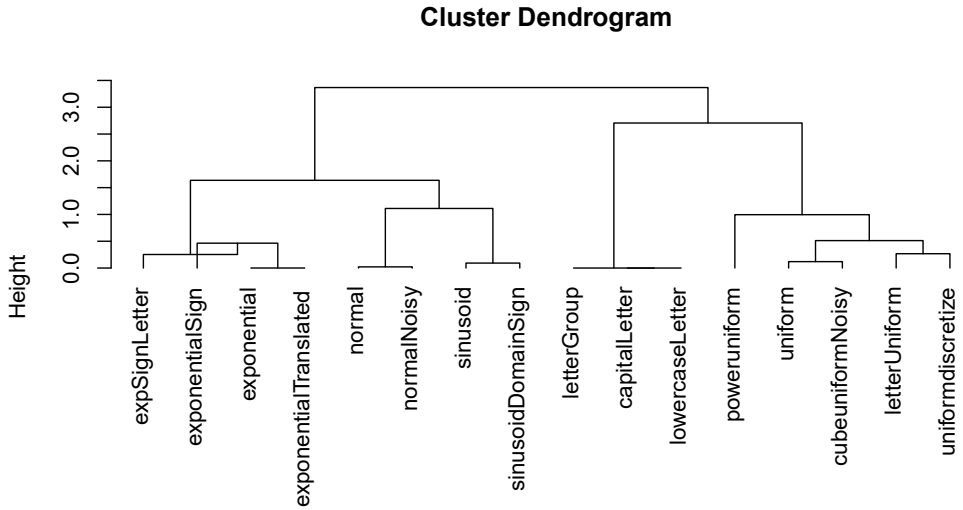


Figure D.2. Clustering of variables for dataset with add of noise $N \sim U(0.01, 0.1)$ using `ClustOfVar`.

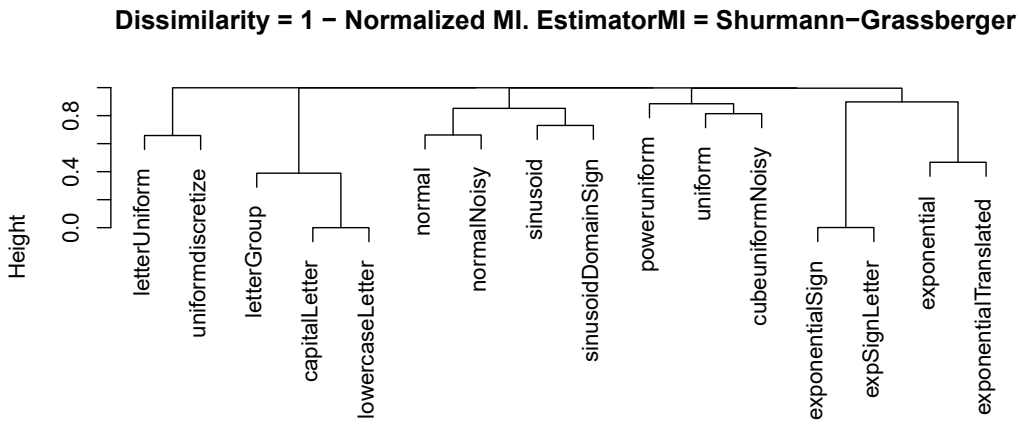


Figure D.3. Clustering of variables for dataset with add of noise $N \sim U(0.11, 0.5)$.

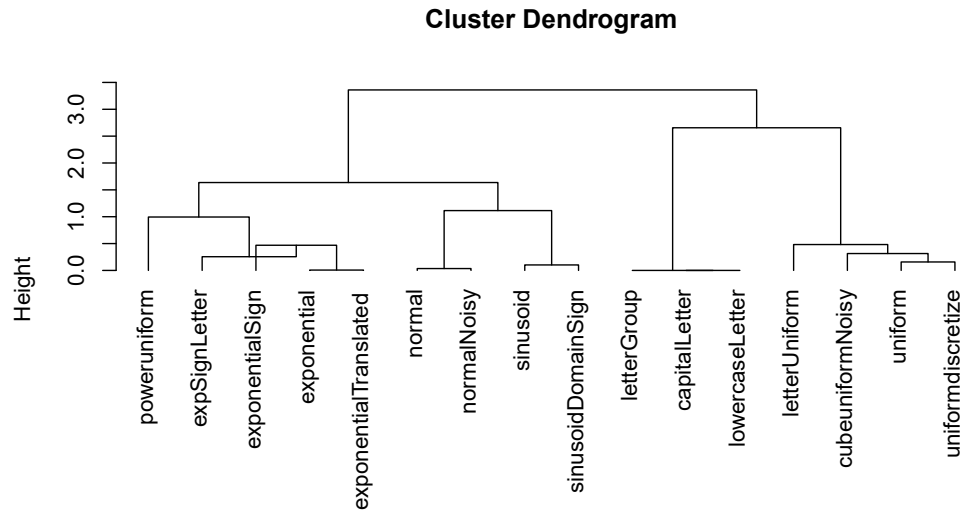


Figure D.4. Clustering of variables for dataset with add of noise $N \sim U(0.11, 0.5)$ using `ClustOfVar`.

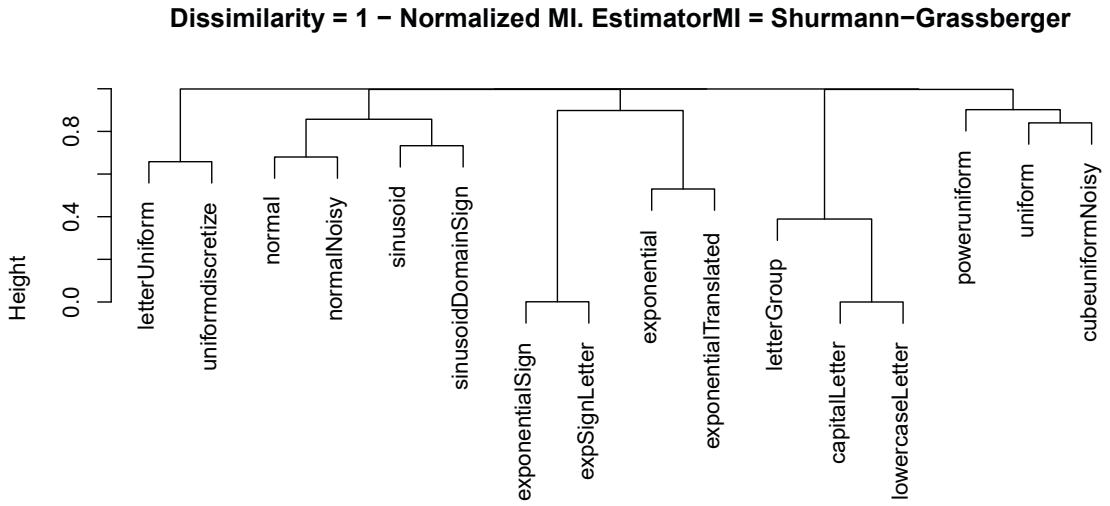


Figure D.5. Clustering of variables for dataset with add of noise $N \sim U(0.51, 1)$.

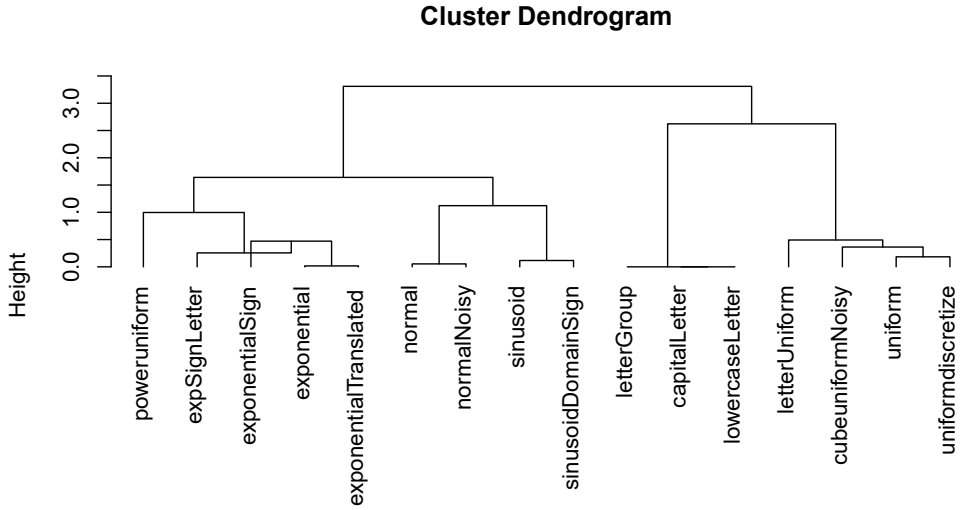


Figure D.6. Clustering of variables for dataset with add of noise $N \sim U(0.51, 1)$ using `ClustOfVar`.

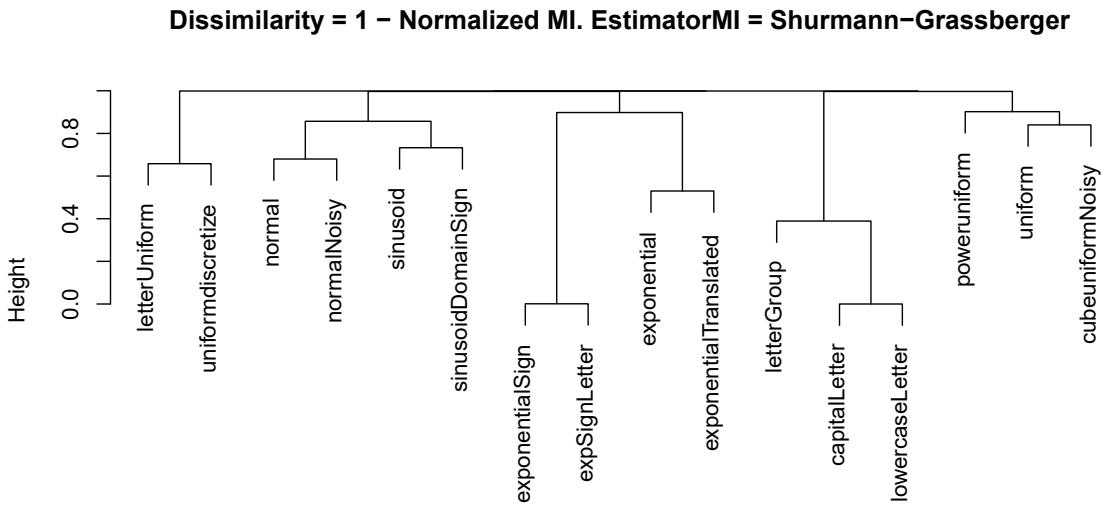


Figure D.7. Clustering of variables for dataset with add of noise $N \sim U(1.01, 1.5)$.

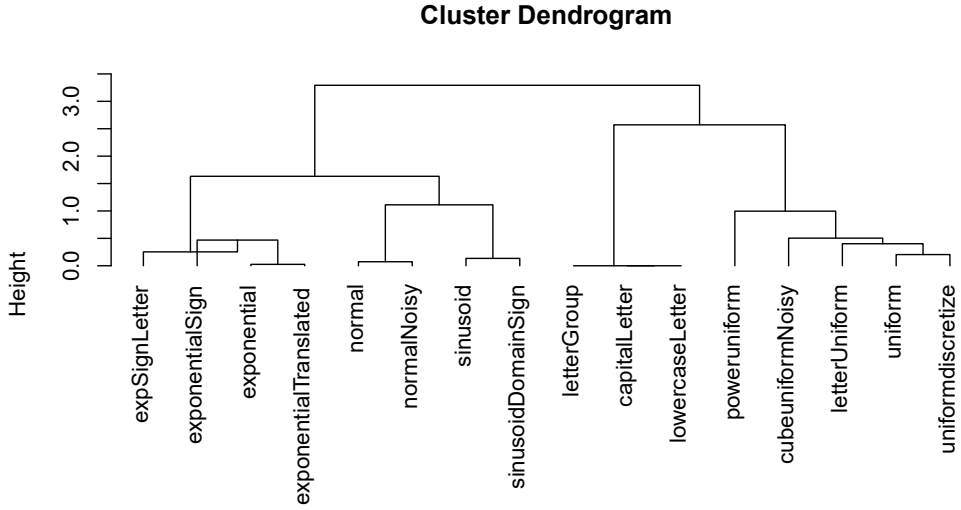


Figure D.8. Clustering of variables for dataset with add of noise $N \sim U(1.01, 1.5)$ using `ClustOfVar`.

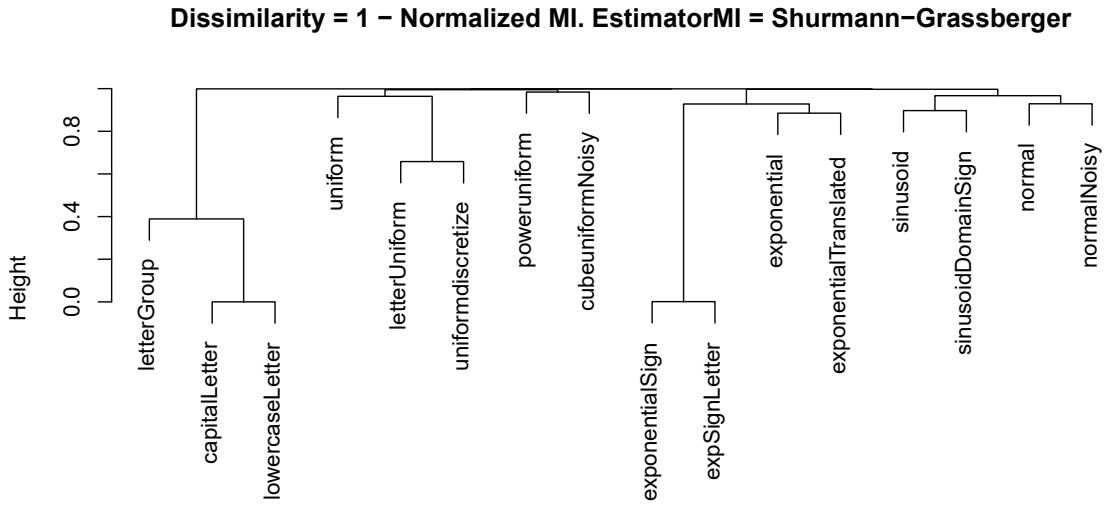


Figure D.9. Clustering of variables for dataset with add of noise $N \sim U(1.5,4)$.

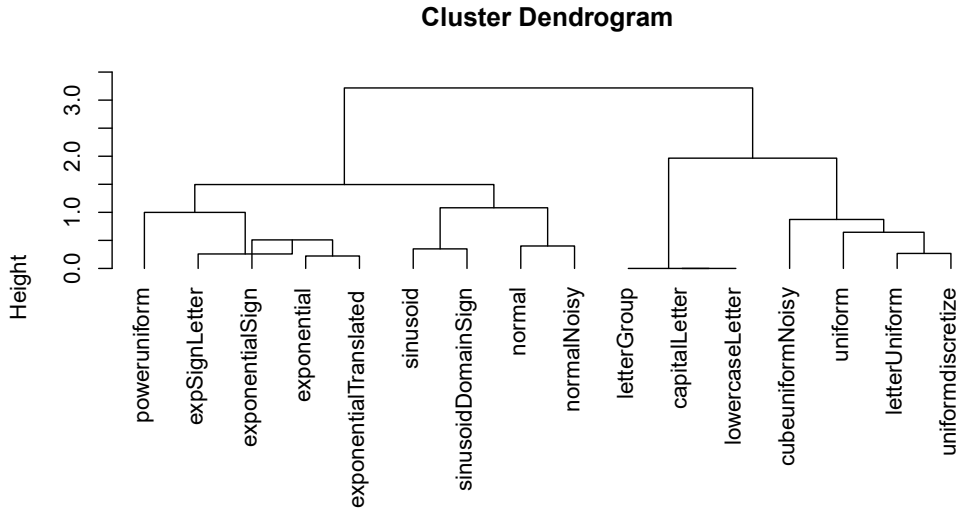


Figure D.10. Clustering of variables for dataset with add of noise $N \sim U(1.5,4)$ using `ClustOfVar`.

Ringraziamenti

Alla conclusione di questo percorso universitario sono doverosi i ringraziamenti. In particolare, ringrazio il Prof. Mauro Gasparini per la fiducia e il supporto datomi durante il corso di laurea. Un sentito ringraziamento al Dr. Pavel Mozgunov per il costante sostegno e i numerosi consigli fornitomi durante la stesura della tesi.

Vorrei ringraziare Luca che mi ha guidato e dato appoggio durante tutto il tirocinio in Tetra Pak. Non da meno sono stati i colleghi del team di APS che ringrazio: Alessandra, Enrico, Gianni, Guido, Luca, Michele, Munib, Sara.

Un ringraziamento speciale va ai miei genitori Stefania e Giuseppe, a mia sorella Federica e a tutta la mia famiglia. Hanno saputo incoraggiarmi e appoggiarmi quotidianamente durante l'intero percorso di studi e hanno sempre creduto in me.

Infine ringrazio tutti i miei amici che hanno allietato i momenti di sconforto.

Bibliography

[1] Shannon C. E. (1948), "A Mathematical Theory of Communication", *The Bell System Technical Journal 27; 3: 379-423*

[2] Kelbert M., Suhov Y. (2013), "Information Theory and Coding by Example", *Cambridge University Press*

[3] Cover M. T., Thomas A. J. (1991), "Elements of information theory", *Wiley series in telecommunications*

[4] Gray R. M. (1990), "Entropy and information theory", *Springer Verlag*, available at: `https://ee.stanford.edu/~gray/it.pdf`

[5] Li W. (1990), "Mutual Information Functions Versus Correlation Functions", *Journal of Statistical Physics 60; 5-6: 823-837*

[6] R Development Core Team (2008), "R: A Language and Environment for Statistical Computing", *ISBN 3-900051-07-0*, available at: `http://www.R-project.org`

[7] Walters-Williams J., Li Y. (2009), "Estimation of Mutual Information: A Survey", *Rough Sets and Knowledge Technology, 389-396*

[8] Rosenblatt M. (1956), "Remarks on Some Nonparametric Estimates of a Density Function", *The Annals of Mathematical Statistics 27; 3: 832-837*

[9] Parzen E. (1962), "On Estimation of a Probability Density Function and Mode", *The Annals of Mathematical Statistics 33; 3: 1065-1076*

[10] Kozachenko L.F., Leonenko N. N. (1987), "On statistical estimation of entropy of random vector", *Problems Information Transmission 23; 2: 95-101*

[11] Leonenko N. N., Pronzato L., Savani V. (2008), "A class of Rényi information estimators for multidimensional densities", *The Annals of Statistics 36; 5: 2153-2182*

[12] Meyer P. E. (2014), "infotheo: Information-Theoretic Measures", *R package version 1.2.0*, available at: `https://CRAN.R-project.org/package=infotheo`

[13] Berrett T. B., Grose D. J., Samworth R. J. (2017), "IndepTest: Nonparametric Independence Tests Based on Entropy Estimation", *R package version 0.1.0*, available at: `https://CRAN.R-project.org/package=IndepTest`

[14] Seok J., Kang Y. S. (2015), "Mutual Information between Discrete variables with Many Categories using Recursive Adaptive Partitioning", *Scientific Reports 5: 10981*

[15] Meyer P. E. (2008), "Information-Theoretic Variable Selection and Network Inference from Microarray Data", *PhD thesis of the Universite Libre de Bruxelles*

[16] Paninski L. (2003), "Estimation of Entropy and Mutual Information", *Neural Computation 15; 6: 1191-1253*

[17] Papana A., Kugiumtzis D. (2008), "Evaluation of mutual information estimators on nonlinear dynamic systems", *Nonlinear Phenomena in Complex Systems 11; 2: 225-232*

[18] Kraskov A., Stögbauer H., Grassberger P. (2004), "Estimating mutual information", *Physical review E 69, 066138*

[19] Gao W., Oh S., Viswanath P. (2017), "Demystifying Fixed k-Nearest Neighbor Information Estimators", *IEEE International Symposium on Information Theory (ISIT)*

[20] Gòmez-Verdejo V., Verleysen M., Fleury J. (2009), "Information-theoretic feature selection for functional data classification", *Neurocomputing 72; 16-18: 3580-3589*

[21] Dawy Z., Hagenuer J., Hanus P., Mueller J. C. (2005), "Mutual Information Based Distance Measures for Classification and Content Recognition with Applications to Genetics", *IEEE International Conference on Communications*

[22] Kraskov A., Stögbauer H., Andrzejak R. G., Grassberger P. (2013), "Hierarchical Clustering Based on Mutual Information", available at: `https://arxiv.org/pdf/q-bio/0311039.pdf`

[23] Linfoot E. H. (1957), "An Informational Measure of Correlation", *Information and control 1; 1: 85-89*

[24] Dionisio A., Menezes R., Mendes D. (2010), "Mutual Information as a Nonlinear Tool for Analyzing Stock Marjet Globalization", available at: `https://pdfs.semanticscholar.org/1fcc/69c987435ef33e0aad72e3cf26ad08202e2c.pdf`

[25] Kojadinovic I. (2002), "Agglomerative hierarchical clustering of continuous variables based on mutual information", *Computational Statistics & Data Analysis 46; 2: 269-294*

[26] Chavent M., Kuentz-Simonet V., Liquet B., Saracco J. (2012), "ClustOfVar: An R package for the Clustering of Variables", *Journal of Statistical Software, 50; 13: 1-16*

[27] Chavent M., Kuentz-Simonet V., Saracco J. (2011), "Orthogonal rotation in PCAMIX", *Advanced in data Analysis and Classification, 6; 2: 131-146*