UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS

Tadeu Junior Gross

Aprendizagem Estrutural de Redes Bayesianas via Perturbação de Dados

São Carlos

2018

Tadeu Junior Gross

Aprendizagem Estrutural de Redes Bayesianas via Perturbação de Dados

Tese apresentada à Escola de Engenharia de São Carlos da Universidade de São Paulo, para obtenção do título de Doutor em Ciências - Programa de Pós-Graduação em Engenharia Elétrica.

Área de concentração: Sistemas Dinâmicos

Orientador: Prof. Dr. Carlos Dias Maciel

São Carlos

2018

Trata-se da versão original.

Ficha catalográfica elaborada pela Biblioteca Prof. Dr. Sérgio Rodrigues Fontes da EESC/USP com os dados inseridos pelo(a) autor(a).

G
878 a

Gross, Tadeu Junior
Aprendizagem Estrutural de Redes Bayesianas via Perturbação de Dados / Tadeu Junior Gross; orientador Carlos Dias Maciel. São Carlos, 2018.

Tese (Doutorado) - Programa de Pós-Graduação em Engenharia Elétrica e Área de Concentração em Sistemas Dinâmicos -- Escola de Engenharia de São Carlos da Universidade de São Paulo, 2018.

1. Rede bayesiana. 2. Média de modelos. 3. Aprendizado de estruturas robustas. 4. Perturbação de dados via bootstrap. 5. Estabilidade de arcos. 6. Limiar analítico. 7. D-separação. 8. Descoberta de Associações. I. Título.

Eduardo Graziosi Silva - CRB - 8/8907

FOLHA DE JULGAMENTO

Candidato: Engenheiro **TADEU JUNIOR GROSS**.

Título da tese: "Aprendizagem estrutural de redes bayesianas via perturbação de dados".

Data da defesa: 29/11/2018.

Comissão Julgadora: **Resultado:**

Prof. Associado **Carlos Dias Maciel**
(Orientador)
(Escola de Engenharia de São Carlos/EESC) APROVADO

Prof. Titular **Jorge Alberto Achcar** APROVADO
(Instituto de Ciências Matemáticas e de Computação/ICMC-USP)

Prof. Dr. **Vinícius Rosa Cota** APROVADO
(Universidade Federal de São João del-Rei/UFSJD)

Prof. Dr. **Ailton Akira Shinoda** APROVADO
(Universidade Estadual Paulista "Júlio de Mesquita Filho"/UNESP – Ilha Solteira)

Prof. Associado **Ricardo Zorzetto Nicoliello Vencio** APROVADO
(Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto/FFCLRP-USP)

Decano do Coordenador do Programa de Pós-Graduação em Engenharia Elétrica:
Prof. Associado **Luís Fernando Costa Alberto**

Presidente da Comissão de Pós-Graduação:
Prof. Associado **Luís Fernando Costa Alberto**

This Thesis is dedicated to my parents
Heitor and Sueli

ACKNOWLEDGEMENTS

First of all, I want to express my more profound gratitude to my parents, Heitor and Sueli, for their unconditional support whenever needed and being my first and eternal encouragers. I also thank both my brother Jonas and my sister Darléia, for being always supportive.

Many thanks to my wife Renata, for her affection, for being an understanding person and for accompanying me on this journey.

I want to thank Prof. Carlos Dias Maciel especially, for his guidance along this doctoral research and also his tolerance during an adverse period.

For all the good times, many thanks to all friends in LPS-EESC-USP, in special to Michel, Jonas, Daniel, Darwin, Jordão, Rafael, Talysson, Matheus and Victor.

Thanks to all friends and colleagues in SEFAZ-MT and POLITEC-MT, in special to João Vicente (*in memory*), Alexis, Adavilso, Elesbão, Fabiano, Festa, Murilo, Emivan, Luiz, Andréia and Willy, for the friendship.

For the support received, I thank the Mato Grosso State Government. In particular, I would like to thank the staff in the personnel department of SESP-MT, for the management of my qualification process.

In a general way, my thanks to everyone who directly or indirectly contributed to this research.

"The task is, not so much to see what no one has yet seen;
but to think what nobody has yet thought,
about that which everybody sees."
Erwin Schrödinger

ABSTRACT

GROSS, T. J. **Structure Learning of Bayesian Networks via Data Perturbation**. 2018. 81p. Tese (Doutorado) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2018.

Structure learning of Bayesian Networks (BNs) is an NP-hard problem, and the use of sub-optimal strategies is essential in domains involving many variables. One of them is to generate multiple approximate structures and then to reduce the ensemble to a representative structure. It is possible to use the occurrence frequency (on the structures ensemble) as the criteria for accepting a dominant directed edge between two nodes and thus obtaining the single structure. In this doctoral research, it was made an analogy with an adapted one-dimensional random-walk for analytically deducing an appropriate decision threshold to such occurrence frequency. The obtained closed-form expression has been validated across benchmark datasets applying the Matthews Correlation Coefficient as the performance metric. In the experiments using a recent medical dataset, the BN resulting from the analytical cutoff-frequency captured the expected associations among nodes and also achieved better prediction performance than the BNs learned with neighbours thresholds to the computed. In literature, the feature accounted along of the perturbed structures has been the edges and not the directed edges (arcs) as in this thesis. That modified strategy still was applied to an elderly dataset to identify potential relationships between variables of medical interest but using an increased threshold instead of the predict by the proposed formula - such prudence is due to the possible social implications of the finding. The motivation behind such an application is that in spite of the proportion of elderly individuals in the population has increased substantially in the last few decades, the risk factors that should be managed in advance to ensure a natural process of mental decline due to ageing remain unknown. In the learned structural model, it was graphically investigated the probabilistic dependence mechanism between two variables of medical interest: the suspected risk factor known as Metabolic Syndrome and the indicator of mental decline referred to as Cognitive Impairment. In this investigation, the concept known in the context of BNs as D-separation has been employed. Results of the carried out study revealed that the dependence between Metabolic Syndrome and Cognitive Variables indeed exists and depends on both Body Mass Index and age.

Keywords: Bayesian Network. Directed Acyclic Graph. Model Averaging. Learning of robust structures. Data perturbation via bootstrap replicas. Stability of arcs. Analytical threshold. Closed-form expression to compute the cutoff-frequency. D-Separation. Associations Discovery. Population Ageing. Cognitive Impairment. Risk Factors. Metabolic Syndrome.

RESUMO

GROSS, T. J. **Aprendizagem Estrutural de Redes Bayesianas via Perturbação de Dados**. 2018. 81p. Tese (Doutorado) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2018.

O aprendizado da estrutura de uma Rede Bayesiana (BN) é um problema NP-difícil, e o uso de estratégias sub-ótimas é essencial em domínios que envolvem muitas variáveis. Uma delas consiste em gerar várias estruturas aproximadas e depois reduzir o conjunto a uma estrutura representativa. É possível usar a frequência de ocorrência (no conjunto de estruturas) como critério para aceitar um arco dominante entre dois nós e assim obter essa estrutura única. Nesta pesquisa de doutorado, foi feita uma analogia com um passeio aleatório unidimensional adaptado para deduzir analiticamente um limiar de decisão apropriado para essa frequência de ocorrência. A expressão de forma fechada obtida foi validada usando bases de dados de referência e aplicando o Coeficiente de Correlação de Matthews como métrica de desempenho. Nos experimentos utilizando dados médicos recentes, a BN resultante da frequência de corte analítica capturou as associações esperadas entre os nós e também obteve melhor desempenho de predição do que as BNs aprendidas com limiares vizinhos ao calculado. Na literatura, a característica contabilizada ao longo das estruturas perturbadas tem sido as arestas e não as arestas direcionadas (arcos) como nesta tese. Essa estratégia modificada ainda foi aplicada a um conjunto de dados de idosos para identificar potenciais relações entre variáveis de interesse médico, mas usando um limiar aumentado em vez do previsto pela fórmula proposta - essa cautela deve-se às possíveis implicações sociais do achado. A motivação por trás dessa aplicação é que, apesar da proporção de idosos na população ter aumentado substancialmente nas últimas décadas, os fatores de risco que devem ser controlados com antecedência para garantir um processo natural de declínio mental devido ao envelhecimento permanecem desconhecidos. No modelo estrutural aprendido, investigou-se graficamente o mecanismo de dependência probabilística entre duas variáveis de interesse médico: o fator de risco suspeito conhecido como Síndrome Metabólica e o indicador de declínio mental denominado Comprometimento Cognitivo. Nessa investigação, empregou-se o conceito conhecido no contexto de BNs como D-separação. Esse estudo revelou que a dependência entre Síndrome Metabólica e Variáveis Cognitivas de fato existe e depende tanto do Índice de Massa Corporal quanto da idade.

Palavras-chave: Rede Bayesiana. Grafo Acíclico Dirigido. Média de Modelos. Aprendizado de estruturas robustas. Perturbação de dados via *bootstrap*. Estabilidade de arcos. Limiar analítico. Expressão fechada para calcular a frequência-de-corte. D-separação. Descoberta de Associações. Envelhecimento da População. Transtorno Cognitivo. Fatores de Risco. Síndrome Metabólica.

LISTA DE FIGURAS

LISTA DE TABELAS

LIST OF ABBREVIATIONS AND ACRONYMS

ACER	Addenbrooke's Cognitive Examination-Revised
BDeu	Bayesian Dirichlet equivalence with uniform prior
BIC	Bayesian Information Criterion
BMI	Body Mass Index
BN	Bayesian Network
CI	Cognitive Impairment
CPD	Conditional Probability Distribution
CPT	Conditional Probability Table
DAG	Directed Acyclic Graph
FN	False Negatives
FP	False Positives
Gly	Glycemia
HC	Hill Climbing
HDL	High-Density Lipoproteins
IC	Inductive Causation
IDF	International Diabetes Federation
JPD	Joint Probability Distribution
KDE	Kernel Density Estimation
K-S	Kolmogorov-Smirnov
MCC	Matthews Correlation Coefficient
MetS	Metabolic Syndrome
ML	Maximum Likelihood
MMSE	Mini-Mental Status Examination
PEBL	Python Environment for Bayesian Learning

PGMPY	Probabilistic Graphical Models in Python
RMS	Root Mean Square
SAH	Systemic Arterial Hypertension
TG	Triglycerides
TN	True Negatives
TP	True Positives
WC	Waist Circumference

LIST OF SYMBOLS

V	Set of nodes
v_i	Node i
B	Bayesian network
n	Number of nodes
G	Structure
E	Adjacency matrix
Θ	Set of parameters
θ_i	Conditional probability distribution to node i
e_{ij}	Element of the adjacency matrix
P_i	Parents of v_i
$\|\theta_i\|$	Amount of probabilities to describe θ_i
$\|\Theta\|$	Number of free parameters
r_i	Number of states to node i
q_i	Number of distinct instantiations of P_i
D	Dataset
N	Number of instances
$\Gamma(.)$	Gamma function
N_{ijk}	Number of instances to which v_i is in k-th state and P_i is in j-th instantiation
K	Number of bootstrap replicas
f_{th}^{+}	Analytical cutoff-frequency
$\mathscr{D}_\ell$	ℓ-th modelled dataset
R	Normalised size of each simulated dataset
MCC	Performance metric

α	Relaxation coefficient
$\mathcal{F}_{th}$	Set of thresholds to which the respective MCCs are close to the maximum value
f_{th}^{*}	Quasi-optimal threshold

SUMÁRIO

1 INTRODUCTION

Data are worthless until one draws knowledge or inferences from them via some statistical model, which can be very simple and rigid, such as a classic linear regression, or sophisticated and flexible, such as a modern Bayesian Network (BN) (GHAHRAMANI, 2015; PUGA; KRZYWINSKI; ALTMAN, 2015). That latter is one of the most effective models in artificial intelligence to reason under uncertainty (YANG et al., 2016), and its data-driven learning has been an issue too exploited lately (YAO; CHOI; DARWICHE, 2017; BARTLETT; CUSSENS, 2017; CAMPOS et al., 2018). This doctoral research concerns the learning of BNs from data.

A Bayesian Network (PEARL, 1988; NEAPOLITAN et al., 2004; KOLLER; FRIEDMAN, 2009) is a concise representation of a joint probability distribution (JPD) that allows the model in a way that is visually transparent, and also aims to capture cause-effect relationship (PEARL, 2009; PEARL; GLYMOUR; JEWELL, 2016). Learning the JPD is a difficult task and modeling it for discrete variables is substantially easier than continuous variables. Also, BNs allow observational inference and causal interventions. A BN consists of a graphical structure augmented by a quantitative part. The graphical structure is a *directed acyclic graph* (DAG) whose nodes and edges express the random variables and probabilistic dependencies among them, respectively. The quantitative part (or CPDs - Conditional Probability Distributions) is the set of all the individual probability distributions, each one attached to a node and conditioned on parents of such node in the DAG. The product of all these conditional distributions provides the JPD (see Equation 2.1).

According to (BIELZA; NAGA, 2014; PUGA; KRZYWINSKI; ALTMAN, 2015), the probabilistic graphical nature of a BN model suggests its use for both *statistical inference tasks*[1] (LEE; YANG; CHO, 2015; ALEXIOU et al., 2017; VIEGAS et al., 2018) and *associations discovery*[2] (VILLANUEVA; MACIEL, 2010). Nowadays, many researchers employ BNs in their practical problems involving uncertainty by exploring at least one of those two abilities - the survey presented in (BIELZA; NAGA, 2014) about BN in neuroscience lists a series of recent articles according to the aim of modeling.

Despite the widely recognized potential and the multiple successful applications, the use of BNs in larger domains (with many variables) remains challenging (ZHAO et al., 2017). One of the main reasons is that the number of candidate networks (DAGs) increases super-exponentially with the number of nodes (ROBINSON, 1977), and finding an optimal

[1] For example, prediction and classification (in (BIELZA; NAGA, 2014), see topics 3, 4.3.1 and 4.3.2).

[2] Uncovering of "latent" relationships among variables from a dataset (see topic 4.2 in (BIELZA; NAGA, 2014)).

directed topology is NP-hard (CHICKERING et al., 1994). Currently, several researchers (KIM; KO; KANG, 2013; VILLANUEVA; MACIEL, 2014; KREIMER; HERMAN, 2016; MADSEN et al., 2017; LIU et al., 2017; CHEN; DARWICHE; CHOI, 2018) have been pursuing sub-optimal strategies for dealing with the NP-hard problem of the structural learning of BNs. Moreover, as with the Chow-Liu approach[3] (CHOW; LIU, 1968; REBANE; PEARL, 1987), much of the available methods focus on inference tasks efficiency[4] (low latency and high accuracy) and provides structures that often do not encode, not even approximately, the real structure of influences among the variables.

A landmark in BN structure learning is the study in (FRIEDMAN; GOLDSZMIDT; WYNER, 1999), which proposed using bootstrap for obtaining more reliable network models. In that approach, multiple learners[5] are used in parallel for providing an ensemble of structures from bootstrap replicas of the data. After that, the ensemble is reduced to a single DAG as follows. In all possible pairs of nodes, the number of times that two nodes have directly linked to each other is counted, and then a threshold is used to decide whether the direct linking indeed exists. Finally, given the linking, its direction is determined by the arc that happened more often.

The above framework provides a DAG containing quite reliable arcs since that they persist to the perturbed versions of the original data (i.e., to the bootstrap replicas). The seminal paper in (FRIEDMAN; GOLDSZMIDT; WYNER, 1999) and the other ones that followed have employed empirical thresholds. In (SCUTARI; NAGARAJAN, 2013), more than a decade later from that seminal work, a systematic way for defining such thresholds was finally proposed. That recent method is statistical-based and has been successfully applied to important studies. However, the procedure in (SCUTARI; NAGARAJAN, 2013) is fully computational and depends on a numerical optimization step. Apropos, no closed-formula to obtain such thresholds was found during the broad literature search conducted before the beginning of the study presented here.

---

[3] Despite not recent, nowadays this technique is still an important alternative for probabilistic modeling of massive datasets (SCHREIBER; NOBLE, 2017), and it is available in computational packages recently developed (e.g., the python package named *pomegranate* (SCHREIBER, 2018)).

[4] This fact is understandable since probabilistic inference using BNs is also NP-hard (in a general way) (COOPER, 1990).

[5] In this context, a learner is some computational algorithm that learns a BN structure for a bootstrap replica of the original data. There are three possible kinds of learners (i.e., BN structure learning algorithms) (SCUTARI; GRAAFLAND; GUTIÉRREZ, 2018): score-based, constraint-based, and hybrid algorithms. Score-based ones (also referred to as search-and-score algorithms) are optimization algorithms that search for a network having a "good" tradeoff between fitness to the data (measured by a score function) and structural complexity. Constrained-based algorithms use conditional independence tests and derive from *Inductive Causation* (IC) algorithm proposed in (VERMA; PEARL, 1991). Finally, hybrid algorithms combine both techniques. For further details, see (SCUTARI; GRAAFLAND; GUTIÉRREZ, 2018).

In this research, the original method of getting a DAG from an ensemble has been modified. Instead of deciding with a threshold if an edge indeed exists (disregarding orientation), here a threshold is directly applied to the most frequent directed edge (arc). One of the principal contributions of this study is the derivation and validation of a closed-formula to that threshold. A comprehensive performance evaluation across simulated data has revealed that the proposed closed-expression provides proper thresholds to obtain a final structure composed only of reliable arcs. In the experiments involving real data, the BN resulting from the analytical cutoff-frequency achieved high success rates in predictions of an essential node, and its directed topology still captured the expected associations among nodes.

The modified approach, based on arcs persistence, also has been applied to model an elderly dataset aiming to uncover some relationship between metabolic syndrome (MetS) and age-related cognitive impairment (CI). Despite the importance of the proposed methodology in this Thesis to the area of modelling under uncertainty, if considered the social aspect, perhaps the practical consequences from such a medical application are even more relevant. In fact, population aging[6] is an inevitable global phenomenon and identifying the risk factors for age-related disorders is becoming an increasingly important research topic (WHO, 2015).

Many researchers (ARAÚJO et al., 2016; CARRILLO et al., 2009; LIVINGSTON et al., 2017) are concerned with finding risk factors for cognitive impairment (CI) - an important predictor of dementia (such as Alzheimer's disease) (ALBERT et al., 2011; NG et al., 2016). Among others, metabolic syndrome (MetS) (ALBERTI; ZIMMET; SHAW, 2005) has been increasingly recognized as a major risk factor for CI in elderly individuals (CHANG; LUNG; YEN, 2015). MetS is a common metabolic abnormality resulting mainly from the combination of genetic propensity and lifestyle factors such as eating habits, sedentary behaviour, and obesity (GRUNDY, 2008).

A review of the literature reveals that the association between MetS and cognitive decline is not completely understood. Some studies (LIU et al., 2015; DEARBORN et al., 2014; FENG et al., 2013; DIK et al., 2007) have suggested that MetS is linked to cognitive decline. Others (ARAÚJO et al., 2017; TOURNOY et al., 2010; KOMULAINEN et al., 2007) found no relationship and, some (LIU et al., 2013; LAUDISIO et al., 2008; BERG et al., 2007) even found a relative improvement in cognition. According to what has been suggested in (CHANG; LUNG; YEN, 2015; EXALTO et al., 2015; WATTS et al., 2013; SIERVO et al., 2014), the controversial findings obtained from similar data collected in different countries indicate that many possibly complex interactions between MetS and cognitive impairment may be present. Thus, further studies to better understand these interactions are still required.

[6] Displacement of the median age to the right.

From a data analysis viewpoint, the studies with different results performed multiple linear (or log-linear) regressions (KRZYWINSKI; ALTMAN, 2015; LEVER; KRZYWINSKI; ALTMAN, 2016) using a dedicated computational package. Although this linear coupling assumption provides a straightforward analysis, its use may not be suitable because knowledge of the underlying process is still lacking and there may be nonlinearities among the predictor variables (see the last paragraph in (KRZYWINSKI; ALTMAN, 2015)).

A natural and entirely appropriate manner for modelling multiple interactions is to build a network of influences (i.e., a causality map or Bayesian Network) (SACHS et al., 2005; KJAERULFF; MADSEN, 2008; FRIEDMAN et al., 2000). This map is a graph G=(V, E), where V and E are compounds of nodes and edges, respectively. The nodes are the measured random variables and edges are the associations between nodes. This network can be computationally learned from the data, knowing its nodes come from a Directed Acyclic Graph (DAG) (NEEDHAM et al., 2007). With this representation, it is possible to search for a direct or indirect link between the variables of interest, and the absence of links may indicate independent behaviour (PEARL, 2009).

According to (PUGA; KRZYWINSKI; ALTMAN, 2015), the identification of latent relationships in large databases is one of the most interesting fields where BNs can be applied. In health studies, the use of a BN can help uncover connections between, for example, diseases and potential (or even unsuspected) risk factors (LUCAS; GAAG; ABU-HANNA, 2004). In the medical field, recent studies employed BNs to extract the dependence between critical variables, for instance: (AUSSEM; MORAIS; CORBEX, 2012) utilized a BN to associate potential risk factors and nasopharyngeal carcinoma and (VEMULAPALLI et al., 2016) used a BN to uncover a non-obvious link between asthma medications and renal failure.

The knowledge codified in a BN is quite dense. In fact, in addition to defining directed edges among nodes (structural learning), the BN learning also involves attaching a probability distribution to each node conditioned to its parents (parameters learning) (NEAPOLITAN et al., 2004). While the BN structure shows the conditional dependencies; its parameters describe the strength of them. Once the BN is fully specified, Bayes' theorem can be used to update the belief (probabilities) about the node states as new information concerning the other nodes is observed (KOLLER; FRIEDMAN, 2009). This Bayesian ability makes the BN a well-suited tool for statistical inference in complex multivariate problems, and its use in the construction of decision support systems is another exciting application possibility (ALEXIOU et al., 2017; BESSANI et al., 2017).

BN modelling is a suitable non-parametric tool for a more comprehensive understanding of the relationship between variables at the same time as the clarifying the MetS-CI dependence in the elderly, a poorly understood medical research issue. Thus, the

suggested BN modelling approach is also employed in this doctoral research to uncover any association between MetS and CI: the learned DAG enabled identification of which conditions imply the (in)dependence of MetS and CI. For this identification task, the D-separation concept (PEARL; GLYMOUR; JEWELL, 2016; HAYDUK et al., 2003) was applied (see section 2.3).

The thesis is organised as follows. Chapter 2 summarises the involved theoretical concepts. Chapter 3 proposes a new strategy for learning robust structures and also an analytical threshold for combining networks learned from bootstrap replicas. Chapter 4 splits into two parts: Part I (section 4.1) describes the collection of the elderly data and the strategy applied to analyse the relationship between MetS and CI on the structure learned via the proposed approach (in this case, an '*ad hoc*' threshold was used); Part II (section 4.2) presents the materials and methods used in the validation of the approach with the analytical threshold. Chapter 5 exposes the results and discussion: its section 5.1 summarises the baseline statistics of the collected dataset, shows that on the learned BN structure there is indeed a path linking MetS to the Cognitive Variables (CI, MMSE-score, and ACER-score[7]), assesses the consistency of the obtained BN structure and explores the dataset properties codified in the graphical representation; its section 5.2 compares the resulting 'good' thresholds to the ones predicted via proposed closed expression and demonstrates that such formula provides efficient cutoff-frequencies indeed. Chapter 6 presents the scientific papers that have been prepared along this doctoral research. Chapter 7 highlights that the dependence between MetS and Cognitive Variables is dependent on both BMI and age, emphasises the potential of the BN-based approach for extracting knowledge from medical data and reinforces that the conducted study demonstrated the validity of both the suggested analytical threshold and also the own proposed learning strategy.

[7] MMSE and ACER refer to the Mini-Mental Status Examination and Addenbrooke's Cognitive Examination-Revised, respectively.

2 FUNDAMENTAL CONCEPTS

In this chapter, BN modelling and structure learning problem are posed formally. Also, essential concepts about search-and-score algorithms, averaging-based structure learning and D-separation are presented. The idea is to provide in this chapter the needed knowledge to understand both the adopted learning context (see section 3.1) and also the tool used to analyse dependencies in an important learned structure (see section 5.1).

2.1 Bayesian networks

A BN model (PEARL, 1988; NEAPOLITAN et al., 2004; KOLLER; FRIEDMAN, 2009) estimates a joint probability distribution (JPD). Formally, a BN for a set $V = \{v_1, v_2, ..., v_n\}$ of n discrete random variables is a pair $B = (G, \Theta)$, where $G = (V, E)$ is a directed acyclic graph (DAG) and $\Theta = \{\theta_1, \theta_2, ..., \theta_n\}$ is a set of conditional probability distributions (CPDs). Usually, the components G and Θ are known as the *structure* and *parameters* of the BN, respectively. In the first component (G), also referred to as *qualitative part*, the vertices (or nodes) are the random variables in V, and the edges, which represent direct dependencies between the variables, are specified in the set[1] $E = \{e_{ij}\}_{i,j=1,...,n}$, where $e_{ij} = 1$ indicates the existence of the arc from v_i to v_j, and $e_{ij} = 0$ means its absence[2]. In the second component (Θ), also referred to as *quantitative part*, the function θ_i corresponds to conditional probability distribution $p(v_i|P_i)$, where P_i are the parents of v_i in the structure[3]. The JPD follows directly from the product of all individual functions in Θ (NEAPOLITAN et al., 2004; KOLLER; FRIEDMAN, 2009):

$$p(v_1, v_2, ..., v_n) = \prod_{i=1}^{n} p(v_i|P_i) \tag{2.1}$$

A data-driven BN learning involves to learn both the *structure* (G) and *parameters* (Θ) from a dataset. The learning of parameters[4] is not addressed in this study, but the number of them needed to specify Θ is employed to normalise the different sizes (number of instances) of simulated data, according explained in chapter 4. See below how to get the number of parameters $|\Theta|$ given the structure.

---

1 The well-known *adjacency matrix* is the set E itself arranged in matrix format.

2 Note that both e_{ij} and e_{ji} should be equal to 0 when the nodes v_i and v_j are not directly connected in the DAG. Also observe that $i = j \Rightarrow e_{ij} = 0$, since the directed graph G should be acyclic.

3 A node v_j is parent of v_i if the BN structure (DAG) has the link $v_j \rightarrow v_i$. In this case, v_i is a child of v_j.

4 Note that probabilities and parameters are interchangeable terms in this context.

2.1.1 Number of free parameters

The following argument is based on (CHICKERING, 1995). Suppose $|\theta_i|$ represents the minimum amount of probabilities to describe $p(v_i|P_i)$ and r_i be the number of states of the discrete variable $v_i \in V$. Then, the number of possible instantiations to the parents P_i of v_i is $q_i = \prod_{v_j \in P_i} r_j$. Also, for every distinct instantiation of P_i, the number of free parameters to the v_i is just $r_i - 1$ since the probabilities sum over all states of v_i should be 1. Therefore, $|\theta_i| = (r_i - 1)q_i$, and the total number of free parameters $|\Theta| = \sum_{i=1}^{n} |\theta_i|$ can be computed by:

$$|\Theta| = \sum_{i=1}^{n} (r_i - 1)q_i. \tag{2.2}$$

2.1.2 Structure learning and scoring functions

The problem of data-driven BN learning can be stated as an optimisation problem: given a dataset D whose N instances (or cases) were drawn from a joint distribution $p(v_1, v_2, ..., v_n)$, find a pair $B^* = (G^*, \Theta^*)$ that maximises the posterior probability of B given D. Formally (BROOM; DO; SUBRAMANIAN, 2012):

$$\begin{aligned} B^* = \underset{B}{argmax}\ P(B|D) &= \underset{B}{argmax}\ P(D|B)P(B), \\ &= \underset{G,\Theta}{argmax}\ P(D|G,\Theta)P(\Theta|G)P(G). \end{aligned} \tag{2.3}$$

In that statement, the term $P(D|G,\Theta)$ is the *likelihood* of the data given the pair $B = (G, \Theta)$, and

$$P(D|G) = \int_{\Theta} P(D|G,\Theta)P(\Theta|G)d\Theta \tag{2.4}$$

is the *marginal likelihood*, which, assuming $P(\Theta|G)$ be a *Dirichlet distribution* with parameters α_{ijk}, can be computed from the data for a given network G using the following expression (HECKERMAN; GEIGER; CHICKERING, 1995):

$$P(D|G) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \tag{2.5}$$

where n, r_i and q_i are the number of nodes, states of the node i and possible instantiations to the parents of the node i, respectively. Additionally, $\Gamma(.)$ is the well-known *Gamma function*, $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$ and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$, where N_{ijk} denotes the number of instances to which both the variable v_i is in its *k-th* state and the set P_i is in its *j-th* possible instantiation[5].

The Bayes' formula implies $P(G|D) \propto P(D|G)P(G)$. Thus, assuming that all DAGs are equally likely (i.e., supposing that $P(G)$ is uniform), a network to which the marginal likelihood in the Equation 2.5 reaches the maximum also maximizes the *a*

[5] Note that P_i represents all parents of v_i in the network G under evaluation.

posteriori probability $P(G|D)$. Therefore, the 'marginalized version' of the optimization problem stated in Equation 2.3, that is

$$G^* = \underset{G}{argmax}\ P(G|D), \tag{2.6}$$

can be solved by searching on the space of DAGs for structures that maximise the expression in Equation 2.5, which in that role of fitness function (or scoring function) is called the *Bayesian Dirichlet equivalence with uniform prior metric* (BDeu). An important property of such metric is the non-distinction among *Markov equivalent structures*: the BDeu scores equally the DAGs that entail the same conditional independencies (BIELZA; NAGA, 2014).

Ideally, a scoring function should provide the exact probability of the data for a given structure (i.e., the value for marginal likelihood shown in Equation 2.4) - the BDeu metric represents an effort in that direction. However, approximations also are often used, like as, for example, the *Bayesian Information Criterion scoring function*, known as BIC. In practice, the approximate metrics are often more efficient, though they are less accurate. Regarding the BIC metric, the approximation is $\ln p(D|G) \approx \ln p(D|\hat{\Theta}_g, G) - \frac{|\hat{\Theta}_g|}{2} \ln N$, where $\hat{\Theta}_g$ and $|\hat{\Theta}_g|$ denote, given G, *the estimated parameters* and *the number of free parameters*[6], respectively; N is the size of the dataset (i.e., the number of instances) (NEEDHAM et al., 2007).

According to the *Occam's Razor Principle*, a BN structure should explain 'well' ('good' fit) and in a simple way ('low' network complexity) the data. The *maximum likelihood* (ML) grows monotonically as the complexity of the structural model increases and in the role of a scoring function would provide a fully connected network. Therefore, the ML is inadequate as a metric for structure learning. The problem would be not only the 'high' complexity of the resulting network. An utterly connected structure does not reflect the essence of the BNs paradigm, in which is the absence of arcs that captures the relevant information of the represented joint distribution (BIELZA; NAGA, 2014; GROSS et al., 2018; NEEDHAM et al., 2007).

The BDeu metric avoids a network with a tangle of unnecessary connections, though its formula is not too explicit as the BIC approximation regarding the penalisation of complexity[7]. Indeed, the BIC clearly prefers the simplest model between the ones that are similarly 'good' because the amount related to complexity (i.e., the piece $\propto |\hat{\Theta}_g|$) is directly subtracted from the part that measures the how well the model explains the data (NEEDHAM et al., 2007).

---

[6] See Equation 2.2 in Section 2.1.1.

[7] A brief discussion about how the BDeu penalises the complexity is given in (SILANDER; KONTKANEN; MYLLYMÄKI, 2007).

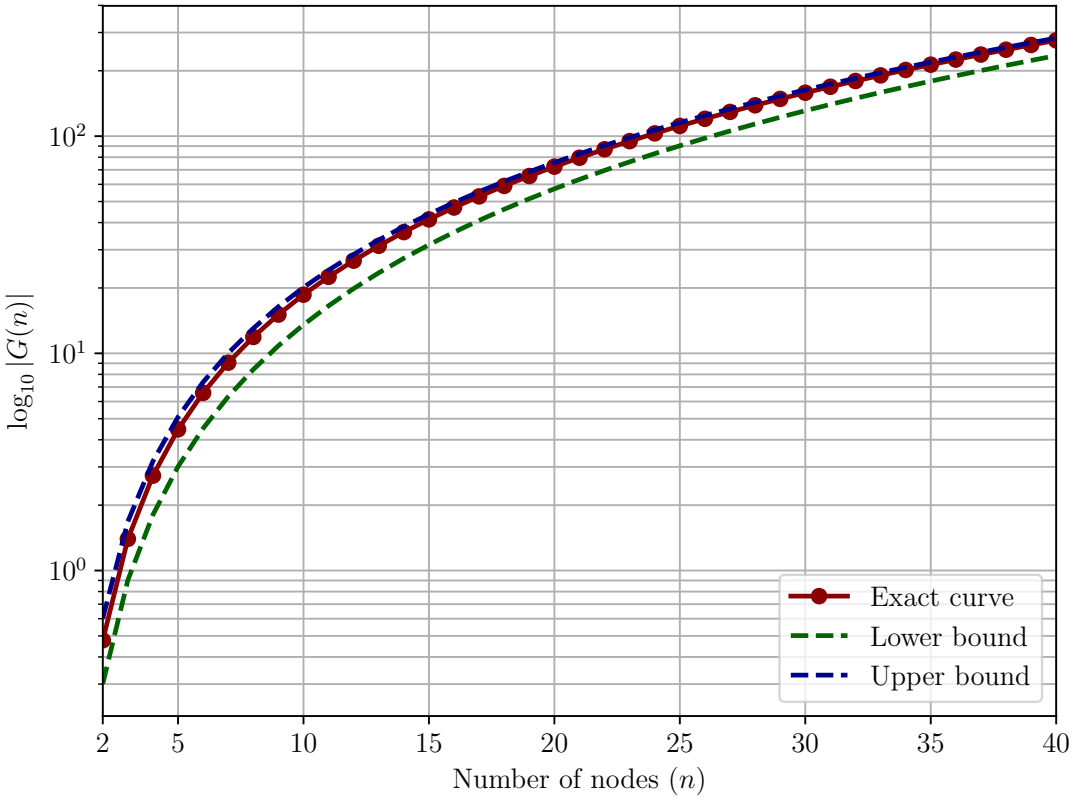
2.1.3 Search-and-score algorithms

In the score-based approach, besides the scoring function, an algorithm is needed for 'walking' over the space of networks in the direction of the structures with the highest scores. Getting to a network that globally maximises a scoring function is a challenge for large domains. The problem is the super-exponential expansion of the search space as the number of nodes increases. The number $|G(n)|$ of possible structures is ever between the bounds $2^{\binom{n}{2}}$ and $n!2^{\binom{n}{2}}$. Such 'combinatorial explosion' of the cardinality is shown in Figure 1, whose exact curve was drawn using the *Robinson's recurrence* (ROBINSON, 1977):

$$|G(n)| = \sum_{m=1}^{n} (-1)^{m+1} \binom{n}{m} 2^{m(n-m)} |G(n-m)|, \tag{2.7}$$

where $|G(n)|$ is the exact number of DAGs for n nodes and $|G(0)| = |G(1)| = 1$.

Figura 1: Number of DAGs as the number of nodes increases according to *Robinson's recurrence*. The lower and upper bounds were computed using the expressions $|G_l(n)| = 2^{n(n-1)/2}$ and $|G_u(n)| = n!2^{n(n-1)/2}$, respectively.



Source: Author

According to Figure 1, the size of space soon becomes prohibitively large for an exhaustive scoring of all networks, and a 'smart' searching algorithm is indeed imperative in score-based structure learning. Currently, the design of efficient search-and-score algorithms is an active and exciting research area (WANG; LIU, 2018; TSIRLIS et al., 2018;

ALONSO-BARBA et al., 2013). Despite the various alternatives, the study in (SCUTARI; NAGARAJAN, 2013) empirically verified that the combination of *Hill Climbing* (HC) algorithm and BDeu metric[8] provides satisfactory networks, above all for utilisation as samples in the averaging approach (see section 2.2).

The greedy heuristic HC traverses the search space by moving between adjacent networks. At each step, neighbour DAGs are visited by adding, deleting or reversing an arc, and the algorithm advances to the one that provides the highest improvement to the scoring function. That procedure is repeated until a local maximum is reached (i.e., the algorithm stops when no neighbour yields improvement to the scoring function). The computational evaluation of neighbour DAGs can be performed efficiently by using a fitness metric that meets the *decomposability property* (e.g., the BDeu can be decomposed as a sum of scoring functions in such a way that only some of them need be updated when an adjacent network is evaluated). Regarding the starting point, the random networks are an option in the case of multiple local searches via HC (GÁMEZ; MATEO; PUERTA, 2011).

2.2 Learning a robust network: the averaging strategy

In exploratory studies, the previous knowledge about the true probability distribution of data is minimum (or absent), and there are no 'gold standards' for directly assessing the learned structures (SCUTARI; NAGARAJAN, 2013). Additionally, real data are commonly 'noisy' and incomplete, and a single run of a search-and-score algorithm can provide a partially spurious structure (FRIEDMAN; GOLDSZMIDT; WYNER, 1999; SCUTARI; NAGARAJAN, 2013; BROOM; DO; SUBRAMANIAN, 2012). For ensuring a reliable network, the study in (SCUTARI; NAGARAJAN, 2013) applied the following steps:

1. For $k = 1, ..., K$:
   a) get a new dataset D_k by resampling (with replacement) the N instances from the original dataset D;
   b) learn the structure $G_k = (V, E_k)$ from D_k using an acknowledged structure learning algorithm (e.g., HC combined to BDeu metric).
2. Estimate the probability that each possible link $v_i - v_j$ is present in true network $G^* = (V, E^*)$ as[9]

$$\bar{e}_{ij} = \bar{e}_{ji} = \frac{1}{K} \sum_{k=1}^{K} (e_{ij}^k + e_{ji}^k), \tag{2.8}$$

[8] Chickering proved in (CHICKERING, 2002) that such combination is consistent.

[9] In Equation 2.8, perceive that the sum $e_{ij}^k + e_{ji}^k$ returns or 0 or 1 for $k = 1, ..., K$ since every adjacency matrix E_k represents a DAG. Also, note that the superscript k is just an index and does not mean a potentiation.

where i, $j \in \{1, ..., n\}$, and e_{ij}^k, $e_{ji}^k \in E_k$.

3. Since there is more belief about links that still are induced when the data are perturbed, interpret $\bar{e}_{ij}$ as the *confidence level* about the link $v_i - v_j$. Then, assume that such link (or edge) is true when $\bar{e}_{ij}$ overcomes a pre-specified threshold f_{th}.

4. For every edge judged as significant, choose as its orientation the direction more frequently observed along the K learned structures. In mathematical terms, fulfil the adjacency matrix E^* as follows[10]:

$$e_{ij}^* = \begin{cases} 0, & \text{if } (\bar{e}_{ij} < f_{th}) \text{ or } (t_{ij} < t_{ji}) \\ 1, & \text{otherwise} \end{cases}, \tag{2.9}$$

where i, $j \in \{1, ..., n\}$ and $t_{ij} = \sum_{k=1}^{K} e_{ij}^k$.

The idea of assessing the edge significance (disregarding orientation) by using *nonparametric bootstrap* (EFRON; TIBSHIRANI, 1994) (step 1) associated to *model averaging* (CLAESKENS; HJORT et al., 2008) (steps 2 and 3) was originally proposed in (FRIEDMAN; GOLDSZMIDT; WYNER, 1999). About accepting the direction observed more frequently as the 'right' orientation of each significant edge (step 4), the research in (SCUTARI; NAGARAJAN, 2013) followed the approach in (IMOTO et al., 2002).

2.3 D-separation

A BN structure visually expresses the circumstances of (in)dependence expressed as algebraic equalities in the probability language, and the D-separation concept can be interpreted as a tool for extracting such circumstances systematically (PEARL; GLYMOUR; JEWELL, 2016; HAYDUK et al., 2003). Given a DAG, two nodes linked through a path (i.e. a sequence of edges disregarding direction) are dependent or D-connected only if none of the basic connections that form such a path are blocked. If this condition is not satisfied, the two focus nodes are independent or D-separated. A basic connection can be (NEEDHAM et al., 2007; PUGA; KRZYWINSKI; ALTMAN, 2015):

- serial ($a \rightarrow b \rightarrow c$): for this type, the chain rule for BNs in Equation 2.1 yields $p(a, b, c) = p(a)p(b|a)p(c|b)$, which, in turn, via marginalization and Bayes' rule, provides $p(c|a) = \sum_b p(b|a)p(c|b)$, $p(a|c) = \sum_b p(b|c)p(a|b)$ and $p(a, c|b) = p(a|b)p(c|b)$. Thus, altering a (or c) affects b and c (or a), but if b is known, c becomes independent of a (conditional independence).

[10] Since E^* is the adjacency matrix for a DAG, it is always true that: $e_{ii}^* = 0$; $(e_{ij}^*).(e_{ji}^*) = 0$; $e_{ij}^* = 1$ implies $v_i \rightarrow v_j$.

- diverging ($d \leftarrow e \rightarrow f$): here, the propagation is similar to that in the above case and evidence of node e blocks the dependence between d and f (conditional independence). The serial and diverging connections are equivalent structures with regard to the joint distribution - for $(a, b, c) = (d, e, f)$, the JPDs are identical.

- converging ($g \rightarrow h \leftarrow i$): in this case, the same chain rule for BNs produces $p(g, h, i) = p(g)p(i)p(h|g, i)$, which, in turn, implies $p(i|g) = p(i)$, $p(i|g, h) \neq p(i)$, and $p(g, i|h) \neq p(g|h)p(i|h)$. Thus, g and i are independent (i.e. the dependence is blocked), but some knowledge of the intermediate variable h (the 'collider node') unblocks the dependence between g and i (conditional dependence). From the sample perspective, the v-structure implies that a scatter plot obtained from values for g and i would not reveal any relationship between these two variables (e.g. a fully symmetric scattering around the origin would be observed). Each value (or range) for h corresponds to a sub-group of points in such a scatter plot - i.e. the cloud of points (g_k, i_k) can be segregated according to h. The collision in h means that it is possible to verify a relationship between g and i only in each sub-group (e.g. an inclined ellipse for each sub-group of the cloud would be observed). In short, the converging structures arise when some data stratification is needed to perceive some dependence between the focus variables that are overall independent.

3 PROPOSED APPROACH

In this chapter, the traditional method of structure learning based on data perturbation is adapted. After that, in this new learning context, an analytical threshold is derived via analogy with a modified one-dimensional random-walk.

3.1 The adaptation

Instead of taking into account the connections confidence (disregarding directions) as in section 2.2, the arcs (directed edges) confidence itself is directly considered. In that adapted approach, the Equation 2.8 has been reduced to the ordinary average of all adjacency matrices, that is,

$$\bar{e}_{ij} = \frac{1}{K}\sum_{k=1}^{K} e_{ij}^{k},\ \ i,j \in \{1,...,n\}, \tag{3.1}$$

and the Equation 2.9, which define the final adjacency matrix E^*, is changed to

$$e_{ij}^{*} = \begin{cases} 0, & \text{if } (\bar{e}_{ij} < f_{th}) \text{ or } (\bar{e}_{ij} < \bar{e}_{ji}) \\ 1, & \text{otherwise} \end{cases},\ \ i,j \in \{1,...,n\}, \tag{3.2}$$

for selecting only the dominant arcs that meet or overcome the cutoff level f_{th}.

The scheme in Figure 2 details the learning approach proposed and adopted along this study. The diagram in 'a' summarises the steps 1 and 2 described in section 2.2, but considering the suggested Equation 3.1 instead of the Equation 2.8. As one can see, the ordinary structures can be obtained in a completely parallel way in practice. The mentioned random initialisations are an attempt to broaden the coverage of the search space by the local learners (i.e., local search algorithms) and achieving an even more robust final network (NAGARAJAN; SCUTARI; LEBRE, 2013). The processing in 'b' generates a mask for discarding the less frequent direction in each link as determined by the Equation 3.2. The procedure in 'c' first carries out an element-by-element multiplication between the average of adjacency matrices and the mask obtained in 'b' - the result is a matrix composed of only arc intensities. After that, still in the line of the Equation 3.2, a threshold is then applied for selecting just the persistent enough arcs, and, in that way, getting the adjacency matrix E^* that specifies the robust structure G^*.

The approach in Figure 2 was first applied to an important medical issue. The concern in this application was to uncover the existence of a strong connection between two specific variables of medical interest and, in this way, an *ad hoc* high cutoff-frequency was adopted to minimise the chance of false arcs (in detriment of false negatives). After that important application, aiming to extend the usability of the suggested approach, a

closed formula to specify more proper thresholds to the arcs was derived. In the following, such a derivation is presented in detail.

3.2 Derivation of an analytical threshold to directed edges

Along of the K learned networks, both spurious and legitimate arcs are present. Intuitively, a robust structure obtained from that ensemble should include just the regular arcs, i.e., those that occur a minimum number of times over such group of networks. The challenge is to set that cuttof level to discard only the arcs that are indeed spurious. To meet it, consider the one-dimensional walk in the Figure 3. In that sketch, the idea is that the interaction between two nodes is influencing the steps in such a way that the observation of the walk adds knowledge about the underlying real connection. The person O_{ij} starts at the point $x_0 = 0$ and, at the position to the instant $k-1$, takes one step whether an arc directly connects the nodes i and j in the network G_k. That step is to the left (-1) or the right $(+1)$ according to the orientation of such arc. Mathematically, the induced walk is described as[1]

$$x_k = x_{k-1} + s_k, \tag{3.3}$$

where $k \in \{1, ..., K\}$ and $s_k \in \{-1, 0, +1\}$. The step is based on $G_k = (V, E_k)$ as follows:

$$s_k = \psi_{ij}(G_k) = e^k_{ij} - e^k_{ji}, \tag{3.4}$$

with $i \in \{1, ..., n-1\}$, $j \in \{i+1, ..., n\}$ and $e^k_{ij}, e^k_{ji} \in E_k$. In this way, $s_k = 0$ or

$$s_k = \begin{cases} -1, \text{if } v_i \leftarrow v_j, \\ +1, \text{if } v_i \rightarrow v_j. \end{cases} \tag{3.5}$$

For a moment, consider the recursion in Equation 3.3 as a stochastic process with $x_0 = 0$ and

$$P(s_k) = \delta(s_k + 1)P_{-1} + \delta(s_k)P_0 + \delta(s_k - 1)P_{+1}, \tag{3.6}$$

where $\delta(.)$ is the well-known *Dirac delta function*. In this way, the distance after many iterations, i.e.,

$$x_K = \sum_{k=1}^{K} s_k, \tag{3.7}$$

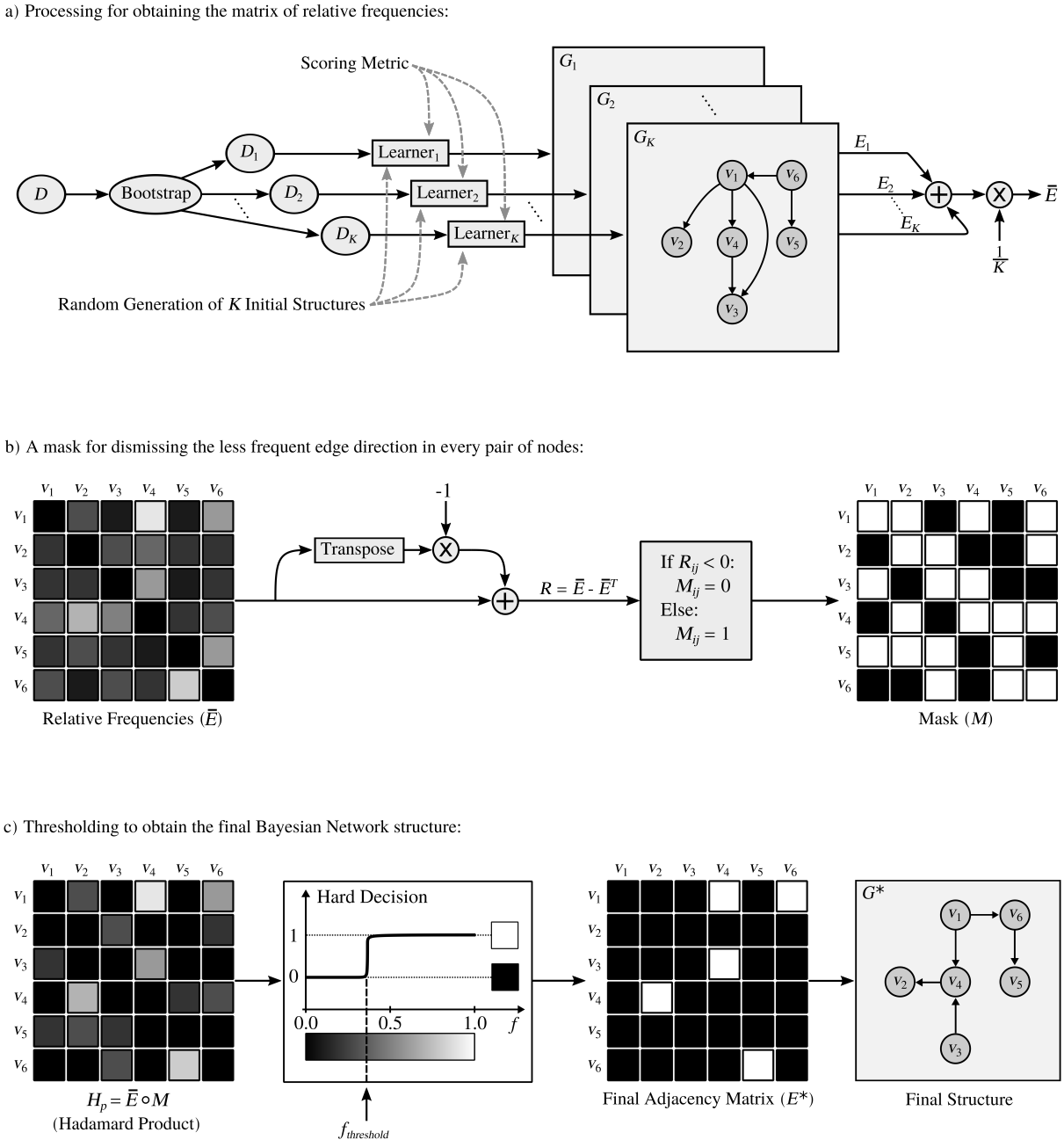
is a random variable with a distribution close to a *Gaussian* curve (it is the *Central Limit Theorem* in action), and only its *mean* and *variance*, both derived below, are enough to characterise it.

The first moment $\overline{x_K} = \mathbb{E}[x_K]$. Given the distribution in Equation 3.6, the operation

$$\overline{x_K} = \sum_{k=1}^{K} \mathbb{E}[s_k] = \sum_{k=1}^{K} [-1.P(-1) + 0.P(0) + 1.P(+1)] \tag{3.8}$$

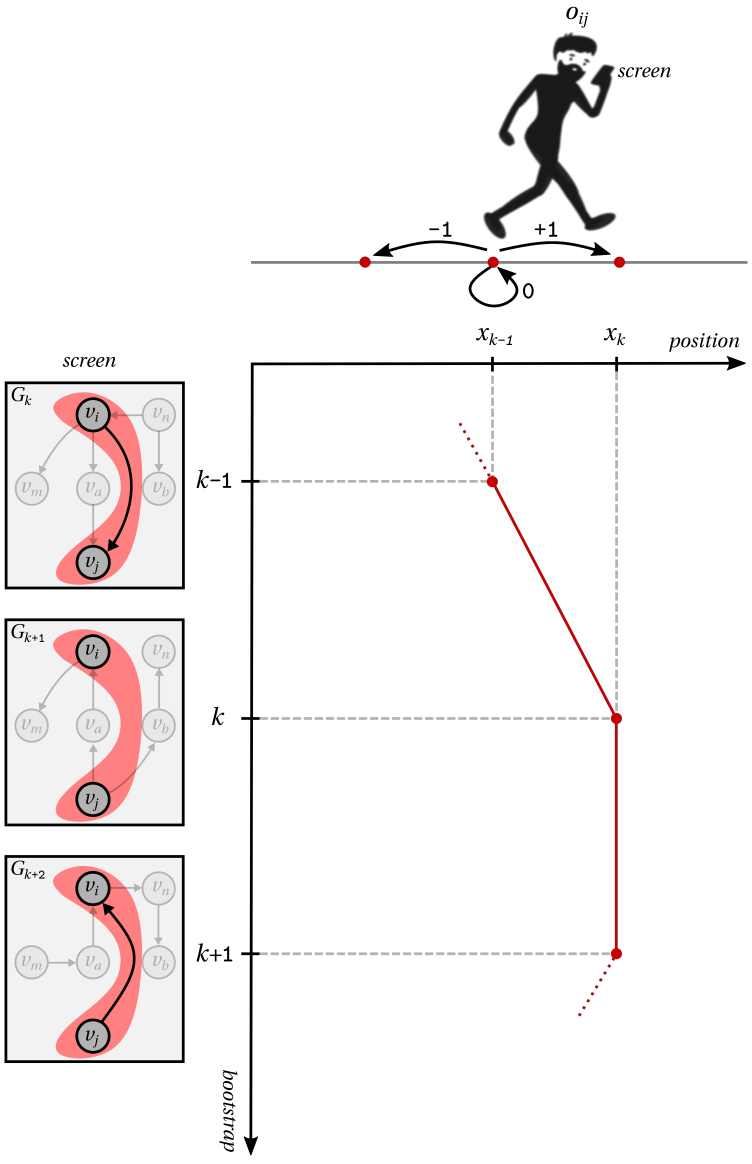
[1] The indices i and j were omitted in x_k and s_k.

Figura 2: A complete scheme for learning a robust network considering the persistence of arcs. This illustration represents the learning context in which the analytical threshold f_{th} (or $f_{threshold}$) is proposed.



Source: Author

Figura 3: A person O_{ij} scrolls a touchscreen to in sequence observe the K learned networks. That observer only pays attention to the interaction between the nodes v_i and v_j along such structures and walks one step whenever observes a direct connection. Such a step is to the left (-1) or right $(+1)$ according to arc orientation of that connection. Consider in detail the sketch itself. In the position x_{k-1}, the walking person updates the screen to network G_k and observes the connection $v_i \to v_j$ (into the red region). Then, the person moves one step to the right and arrives at $x_k = x_{k-1} + 1$. In this new position, O_{ij} scrolls to the structure G_{k+1} and does not see a direct connection between the focus variables (the red area is empty). In this case, O_{ij} remains in the same position (i.e., $x_{k+1} = x_k$) until the next observation. The person conduct says a lot about the connection between v_i and v_j. A 'drunkard' behaviour, i.e., a completely erratic walking around the starting position $x_0 = 0$, increases the belief that the local changes (i.e., -1, 0, 1) have similar probabilities and that the randomness is dominating the emission of the arcs to the focus connection. On the other hand, as more the person moves away from the starting point in one direction, higher the belief that a specific arc orientation is predominating in the emissions.



Source: Author

provides

$$\overline{x_K} = (P_{+1} - P_{-1}).K. \tag{3.9}$$

The second moment $\overline{x_K^2} = \mathbb{E}[x_K^2]$. Considering a realisation $\{x_0, x_1, ..., x_k, ..., x_{K-1}, x_K\}$, the randomness of s_K imposes three possibilities to x_K: $x_{K-1} - 1$, x_{K-1} and $x_{K-1} + 1$. Consequently,

$$x_K^2 = (x_{K-1} - 1)^2 \text{ or } (x_{K-1})^2 \text{ or } (x_{K-1} + 1)^2. \tag{3.10}$$

In a 'large' number of independent realisations, the components in Equation 3.10 should occur in the same proportion of their respective probabilities, which in turn are specified in Equation 3.6. In this manner,

$$\begin{aligned} \mathbb{E}_s\left[x_K^2\right] &= P(-1).(x_{K-1} - 1)^2 + P(0).(x_{K-1})^2 + P(+1).(x_{K-1} + 1)^2 \\ &= x_{K-1}^2 + 2.(P_{+1} - P_{-1}).x_{K-1} + (P_{+1} + P_{-1}), \end{aligned} \tag{3.11}$$

and $\overline{x_K^2} = \mathbb{E}_x\left\{\mathbb{E}_s\left[x_K^2\right]\right\}$ is

$$\overline{x_K^2} = \overline{x_{K-1}^2} + 2.(P_{+1} - P_{-1}).\overline{x_{K-1}} + (P_{+1} + P_{-1}). \tag{3.12}$$

The mean in Equation 3.9 implies $\overline{x_{K-1}} = (P_{+1} - P_{-1}).(K-1)$, then the Equation 3.12 can be written as:

$$y_K = y_{K-1} + 2.D^2.(K-1) + S, \tag{3.13}$$

where $y_K = \overline{x_K^2}$, $D = (P_{+1} - P_{-1})$ and $S = (P_{+1} + P_{-1})$. The sum from 1 to K on both sides of that equation, i.e.,

$$\sum_{k=1}^{K} y_k = \sum_{k=1}^{K} y_{k-1} + \sum_{k=1}^{K} 2.D^2.(k-1) + \sum_{k=1}^{K} S, \tag{3.14}$$

provides

$$y_K = y_0 + 2.D^2.\sum_{k=1}^{K}(k-1) + K.S, \tag{3.15}$$

which in turn is equal to

$$y_K = K.(K-1).D^2 + K.S, \tag{3.16}$$

since that $y_0 = 0$ and $\sum_{k=1}^{K}(k-1) = K.(K-1)/2$. Therefore, the second moment to the distance achieved after K iterations is

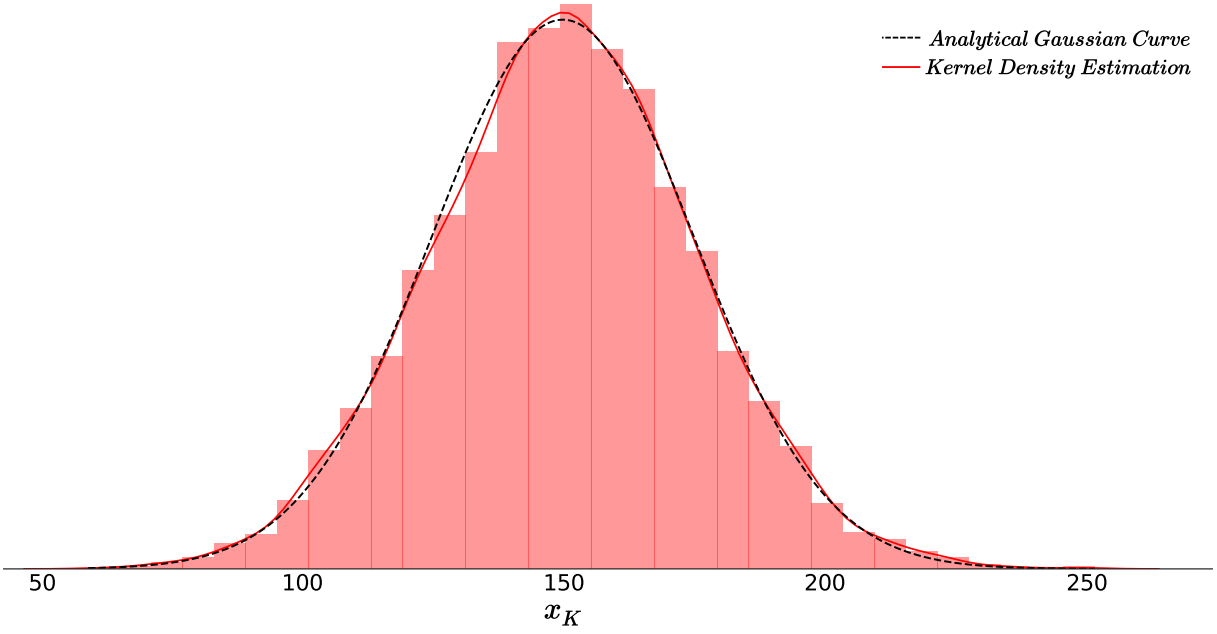
$$\overline{x_K^2} = K.(K-1).(P_{+1} - P_{-1})^2 + K.(P_{+1} + P_{-1}). \tag{3.17}$$

The variance $\sigma_{x_K}^2 = \mathbb{E}\left[(x_K - \overline{x_K})^2\right]$. Considering the moments in Equations 3.9 and 3.17 and the statistical identity $\sigma_{x_K}^2 = \overline{x_K^2} - \overline{x_K}^2$, the following expression to the variance of x_K is obtained:

$$\begin{aligned} \sigma_{x_K}^2 &= K.(S - D^2) \\ &= K.\left[(P_{+1} + P_{-1}) - (P_{+1} - P_{-1})^2\right] \end{aligned} \tag{3.18}$$

The Figure 4 shows the histogram to the distance x_K at instant $K = 1000$ considering 10000 realisations of the idealised stochastic process. In that numerical test, $P_{-1} = 0.25$, $P_{+1} = 0.40$ and $P_0 = 0.35$ were the adopted probabilities to every s_k, $k = 1, ..., K$. The parameters pair $(\overline{x_K}, \sigma^2_{x_K})$, calculated according to the deduced Equations 3.9 and 3.18, defines the black dashed Gaussian[2]. As can be seen, that analytical curve fitted 'well' to histogram and 'almost coincided' to *KDE* profile[3].

Figura 4: An histogram and a *KDE* curve to x_K. As can be seen, the simulation values are very well represented by the Gaussian curve with the parameters calculated through the formulas deduced in this section.



Source: Author

The outlined stochastic interpretation is useful for assessing the induced walk nature and concluding about the underlying network connection. The absence of any trend regarding the kind of interaction between the two focus nodes means that $P_{-1} = P_{+1} = P_0 = 1/3$ and produces an utterly erratic movement around the origin $x_0 = 0$. According to the developed formulas, namely the Equations 3.9, 3.17 and 3.18, such a random-walk does not present a drift phenomenon and, after K steps, the distance has *zero mean* and mean square value, as well as variance, given by

$$\overline{x_K^2} = \sigma^2_{x_K} = (2/3).K. \tag{3.19}$$

In this way, the '*effective distance*' (or '*rms*[4] *distance*') achieved by the resulting

[2] The developed formulas provided $(150.0, 627.5)$. The simulated values yielded a close pair, namely $(150.2, 624.2)$.

[3] *KDE* means *Kernel Density Estimation*.

[4] The acronym '*rms*' means '*root mean square*'.

random-walk is

$$x_{rms} = \sqrt{(2/3).K}, \tag{3.20}$$

and K^+, the sum of steps to the right, has the following '*effective deviation*' from its expected value $K/3$:

$$\left(K^+ - K/3\right)_{rms} = x_{rms}/\sqrt{3} = (1/3).\sqrt{2.K}. \tag{3.21}$$

That expression implies that the proportion f^+ of unitary displacements to the right in a realisation $\{x_k\}_{k=0}^{k=K}$ is

$$f^+ = (K^+/K) \pm (1/3).\sqrt{2/K}. \tag{3.22}$$

Therefore, when none kind of connection of the triplet[5] predominates between the two focus nodes along of the K learned structures, the relative frequency f^+ tend to $1/3$ with a standard deviation

$$\sigma_{f^+} = (1/3).\sqrt{2/K}. \tag{3.23}$$

A realisation of x_k with a f^+ that moves away from its expected value beyond $3.\sigma_{f^+}$ perhaps be a product from 'pure' randomness, but the chance of this is minimal. In such a case, it sounds wiser to admit an underlying regularity. With this in mind, it seems reasonable not to ignore the most persistent arc between two nodes when its frequency is higher than

$$f^+_{th} = (1/3) + 3.\sigma_{f^+} = (1/3) + \sqrt{2/K} \tag{3.24}$$

That is the suggested cutoff frequency for discarding the spurious arcs on the learned ensemble and thus ensuring a reliable structure. Although the strict optimal threshold may depend on both the data and the adopted structure learning algorithm (SCUTARI; NAGARAJAN, 2013), the experiments showed that the cutoff level obtained via such a formula is a quite reasonable choice in practice.

[5] '$\rightarrow$', '$\leftarrow$' and '*unconnected*'.

4 METHODS

The adapted approach (section 3.1) firstly was applied to analyse an essential and current medical issue - the problem nature itself allowed the use of a quite conservative threshold[1] and, in this way, an early application of the procedure in Figure 2, i.e., even before derivation of the formula to compute a more proper cutoff-frequency (section 3.2). After that medical application, the experimental validation of the proposed analytical threshold was accomplished. Therefore, the study as a whole has been split into two parts, each one of them with its particular methodology.

4.1 Part I: Methodology used in the real application of the adapted approach

The application part consisted of the modelling of a recent dataset obtained via a partnership to *Cognitive-behavioural Neurology Service* (CNS) of Federal University of São Carlos (UFSCar), Brazil. In the following, all involved procedures are described and justified.

4.1.1 Data Collection

In obedience to Resolution 466/12 of the National Health Council in Brazil, the Project was submitted to the Human Research Ethics Committee of the Federal University of São Carlos, and data collection started after its approval by Technical Opinion 1.680.626.

In this cross-sectional study, two hundred elderly participants aged at least 65 years were interviewed and assessed at ten Public Healthcare Units[2] (20 elderly participants per unit) uniformly distributed throughout São Carlos, Brazil, between August 2016 and April 2017. The sample size of 200 expresses the maximum number of elderly participants (instances) recruited in this period[3] with the available resources. Although small, the sample size is similar to those employed in some comparable studies (LIU et al., 2015; KOMULAINEN et al., 2007; LIU et al., 2013; LAUDISIO et al., 2008; BERG et al., 2007).

One cannot expect to obtain a detailed model from such few instances with so many attributes (measured variables). However, the collected dataset encloses valuable information as to the tendency for associations between the domain variables. According to (KOLLER; FRIEDMAN, 2009)[4], to obtain a reliable structure of linkages between the nodes in the face of data scarcity, one can learn several high-score structures and subsequently merge them. In this study, this strategy was followed (see subsection 4.1.3).

1 The employ of an undoubtedly high cutoff-frequency.
2 Referred to as UBS or USF in Brazil.
3 The data collection occurred during a short master's degree program.
4 See highlights on pages 825 and 1043 of (KOLLER; FRIEDMAN, 2009).

At the included Healthcare Units, any older adult who presented during the day was invited to take part in the study. Every included participant signed an informed consent form. Individuals previously diagnosed with depression or with a history of brain tumour or stroke were not invited to join the study. Measured variables include demographic and anthropometric features, blood indicators of metabolic syndrome, and the cognitive decline score.

The above data gathering procedure is a simple, fast, and cheap and is known as convenience sampling (POLIT; BECK, 2010). In exploratory studies, the use of available instances from a population has been the norm instead of the exception (JAGER; PUTNICK; BORNSTEIN, 2017). In fact, (BORNSTEIN; JAGER; PUTNICK, 2013) verified that in five prominent journals, 92.5% of their publications from 2007 to 2011 employed convenience sampling[5]. While the adopted strategy may not be optimal, it is commonly used in the scientific literature.

4.1.1.1 Demographic and Anthropometric Variables

The recorded features included the age (years), formal education (years), weight (kg), height (m), and waist circumference (cm) of all participants. Categorical features such as systemic arterial hypertension (SAH)[6], sex, physical activity, pharmaceutical drug user, alcoholic consumption, and cigarette smoking.

4.1.1.2 Biological Samples

A blood sample was extracted from each participant after at least 12 hours of fasting. These samples were submitted to a laboratory in order to measure the levels in mg/dL of triglycerides (TG), glycemia (Gly), and high-density lipoproteins (HDL). Here, TG, Gly, and HDL are binary variables used for classifying the correspondent laboratory results as either 'ordinary-0' or 'altered-1' according to the following guidelines: 1) TG = 1 when the triglycerides level was $\geq$150 mg/dL or the participant was under specific treatment due to a previous augmented level; 2) Gly = 1 when the glycemia level was >100 mg/dL or there was a previous diagnosis of Type 2 Diabetes; 3) HDL = 1 when the high-density lipoproteins level was $< L_{th}$ or if the participant was under specific treatment due to a previous low-level. The threshold L_{th} was set at 40 mg/dL for men and 50 mg/dL for women.

[5] See page 3 of (JAGER; PUTNICK; BORNSTEIN, 2017).

[6] SAH were considered present when systolic and diastolic pressures were greater than 130 and 85 mmHg, respectively; or if the subject was being treated for hypertension.

4.1.1.3 Metabolic Syndrome Variable - MetS

The above guidelines for SAH, TG, Gly, and HDL are criteria considered by the International Diabetes Federation[7] (IDF) in the definition of Metabolic Syndrome (MetS). According to the IDF, an obese adult has this syndrome (i.e. MetS = 1) when at least two of those criteria are present (i.e. whether SAH + TG + Gly + HDL $\geq$ 2) (ALBERTI; ZIMMET; SHAW, 2005). An adult is considered obese by the IDF when at least one of the two following conditions is met: 1) body mass index (BMI) $>$ 30 kg/m^2; 2) Waist Circumference (WC) $>$ 90 cm for men and $>$ 80 cm for women.

4.1.1.4 Cognitive Impairment Variable - CI

The ACE-R protocol (MIOSHI et al., 2006; CARVALHO; BARBOSA; CARAMELLI, 2010) was applied to all participants. The cognitive level of each participant was then obtained using the ACER and MMSE scales[8]. To decide whether cognitive impairment was present (i.e. CI = 1) taking into account schooling, the obtained MMSE score was compared with the threshold chosen according to (BRUCKI et al., 2003).

4.1.2 Dataset Format

After the data collection, the continuous variables were discretized and the dataset actually used for learning the discrete BN was the spreadsheet of 200 rows (200 elderly participants) containing the following 19 measurements (19 columns)[9]: **Age** (65 to 90 years, 9 bins), **Gender** (0-woman or 1-man), **Educ** (formal education, 0 to 15 years, 15 bins), **Exercise** (regular physical activity, 0-no or 1-yes), **Drug Use** (0-no or 1-yes for the use of some pharmaceutical drugs), **Smoking** (0-no or 1-yes), **AlcBev** (Alcohol Consumption, 0-no or 1-yes), **Weight** (44 to 122 kg, 15 bins), **Height** (1.42 to 1.78 m, 11 bins), **BMI** (Body Mass Index, 18 to 49 kg/m^2, 15 bins), **WC** (Waist Circumference, 60 to 130 cm, 11 bins), **TG** (Triglycerides, 0-ordinary or 1-altered), **Gly** (Glycemia, 0-normal or 1-abnormal), **HDL** (High-Density Lipoproteins, 0-normal or 1-abnormal), **SAH** (Systemic Arterial Hypertension, 0-not present or 1-present), **MetS** (Metabolic Syndrome, 0-not present or 1-present), **ACER** (Addenbrooke's Cognitive Examination Revised, 0 to 100, 11 bins), **MMSE** (Mini-Mental Status Examination, 0 to 30, 15 bins) and **CI** (Cognitive Impairment, 0-not present or 1-present).

For each binned variable, the number of bins was calculated using the '*auto*' option of the '*numpy.histogram*' function[10]. This method combines the 'Sturges'[11] (STURGES,

[7] https://www.idf.org

[8] The MMSE score was obtained directly from the ACE-R protocol results

[9] In parentheses, there is a brief description of the respective variables. For continuous variables, the support of the probabilities and the number of bins used for uniform discretization is indicated. For categorical variables, the meaning of each numeric label is stated.

[10] https://docs.scipy.org/doc/numpy/reference/generated/numpy.histogram.html

[11] Default method in R.

1926) and 'Freedman Diaconis' (FREEDMAN; DIACONIS, 1981) estimators and results for sufficient all-around performance (OLIPHANT, 2015).

4.1.3 Data Analysis Strategy

The dependencies between the variables were captured through a fully graphical and non-parametric strategy; the influences network was represented through a BN structure learned from the discretized dataset. Since learning the DAG would be too complicated or even impossible for a human (VILLANUEVA; MACIEL, 2014), this task was performed using the greedy search algorithm of the PEBL[12] (SHAH; WOOLF, 2009).

In the experiment, an ensemble was first generated with $K = 1000$ DAGs by running the greedy search algorithm one thousand times. The different data used in each one of these runs were generated 'perturbing' the original data. The underlying idea is that there is less uncertainty regarding the arcs that are still induced when the data are 'perturbed' (FRIEDMAN; GOLDSZMIDT; WYNER, 1999). The diversity of structures is due to both data 'perturbation' (190 out of 200 elderly in participants were randomly selected for each complete run of the greedy search algorithm)[13] and the greedy search process itself (its initialization is always random and the local optimizations performed during a run are also nondeterministic). One million iterations were used as the stopping criterion for each complete run of the greedy algorithm. After that, the ensemble was reduced to a single 'consensus' DAG (model-averaging approach). In this reduction, the number of times with which each one of the three possible connections (i.e., '$\leftarrow$', '$\rightarrow$', and 'absent') occurred between every pair of nodes in the obtained 1000 graphs was counted. Only the directed edges that occurred at least 67% of the time were accepted.

In an exploratory study, greedy hill-climbing with random restarts seemed to be an adequate technique because it has few optimization parameters to set (here, only the total number of iterations) and was already used in a similar model-averaging approach (FRIEDMAN; GOLDSZMIDT; WYNER, 1999). As for the stopping criterion, 10^6 iterations were sufficient to reach a stable solution (local optimum) in each complete run. Regarding the ensemble size, the choice of $K = 1000$ resulted in a tolerable runtime and appeared appropriate because the similar approach in (FRIEDMAN; GOLDSZMIDT; WYNER, 1999) that used a lower value. Concerning data 'perturbation', the small value of 5% (random withdrawal of 10 rows) was the adopted method of 'perturbing' the dataset without affecting its general statistical properties significantly.

Regarding the threshold, the 67% level was a simple way of ensuring high reliability of obtaining pathways between the nodes (in this medical application, to accept a false

[12] Python Environment for Bayesian Learning.

[13] Although such a data perturbation strategy is not precisely the 'nonparametric bootstrap' mentioned in section 2.2, a posterior test has revealed an equivalent result regarding the learned structure.

linkage is worse than rejecting a true one). With that value, every arc in the final network necessarily occurred two times more than its complement[14]. The use of a minimum occurrence rate slightly higher than 33.33% to legitimise the most frequent arc of the triplet would imply to reject the two remain arcs with occurrence rates of approximately (i.e. slightly lower than) 33.33% for each one. This closeness between the acceptance and rejection levels would affect the credibility of an eventual linking path between MetS and CI in the resulting DAG.

In the learned DAG, the nodes that have no interconnection (direct or indirect) path between them are considered independent (PEARL, 2009). Thus, since it is aimed to primarily show the association between metabolic syndrome (MetS) and cognitive impairment (CI), the final representation of the DAG was even more simplified by excluding all nodes that remained isolated from these two focus nodes. From this compact DAG, it was possible to infer every conditional (in)dependence between MetS and CI variables by analysing, with the D-separation concept in mind (see Section 2.3), the linking path between the two correspondent nodes.

The obtained BN model (structure) as well as its analysis via D-separation are matters treated in the next chapter. In the following section, the concern is to describe the adopted conduct to validate the threshold proposed in section 3.2.

4.2 Part II: Methodology employed in the validation of the proposed threshold

This section describes the methodology used to verify if the threshold in Equation 3.24 is a proper choice in the learning context sketched in Figure 2. The modelled datasets and whole experimental setup are presented in detail, as well as the strategy to evaluate of resulting networks and the computational resources employed.

4.2.1 Modelled datasets

A convenient way to verify the performance of a structure learning procedure is directly to confront the resulting network against the correct one. In that common strategy (SCUTARI; NAGARAJAN, 2013; BROOM; DO; SUBRAMANIAN, 2012), the true structure itself is used to generate simulated data which in turn are modelled using the learning process under test. In the real world, the right structure model to collected data is unknown in general. In that case, previous knowledge about some variables of the domain can be utilised to inspect pieces of the learned structure. The intuition is that, if such expected local relationships are being captured, the structure learning process as a whole is performing well. In this experiment, a recently collected dataset and three simulated ones are used in the tests, and all are briefly described below.

[14] The other two 'arcs' of the triplet.

Asia ($\mathscr{D}_1$). This dataset is obtained via simulation of the influences network to Dyspnea described in (LAURITZEN; SPIEGELHALTER, 1988). In that network, some lung diseases are possible direct causes to such a breathing disorder, and the visit to Asia is a risk factor to one of them. Such a causal network has $n = 8$ nodes, 8 arcs and $|\Theta|_1 = 18$ parameters (see Section 2.1.1).

Survey ($\mathscr{D}_2$). This dataset is gotten by simulating the causal network idealised in (SCUTARI; DENIS, 2014). In the conception of that network, (SCUTARI; DENIS, 2014) considered a hypothetical survey about the influence of demographic and socioeconomic features[15] on the means of transport chosen by the people. That fictitious network has $n = 6$ nodes, 6 arcs and $|\Theta|_2 = 21$ parameters.

Sachs ($\mathscr{D}_3$). Unlike previous two datasets, this one is not synthetic. This experimental dataset contains the biochemical measurements on single cells that were used in Sachs et al. (SACHS et al., 2005) for uncovering causal relationships in cellular signalling networks. In this case, the structure broadly acknowledged as true has $n = 11$ nodes, 17 arcs and $|\Theta|_3 = 178$ parameters.

In practice, the three networks characterised above were simulated to generate new datasets (referred to as $\mathscr{D}_{\ell,R}$), each one of them with a real size (number of instances) given by

$$N_{\ell,R} = R.|\Theta|_\ell, \;\; R \in \mathscr{R}, \tag{4.1}$$

where $\mathscr{R} = \{0.25, 0.5, 0.75, 1, 2, 5, 10, 15, 20, 25\}$ is the set of the chosen normalised sizes. Observe that the amount of free parameters $|\Theta|_\ell$ is not equal across the different true networks, and it seems reasonable to use larger datasets to learn the structures with more unknown parameters. A similar strategy of normalisation was used in (SCUTARI; NAGARAJAN, 2013).

Due to the limited hardware resources and to prevent an excessive runtime, the benchmark networks with more than 11 nodes, as those used in (SCUTARI; NAGARAJAN, 2013) and (BROOM; DO; SUBRAMANIAN, 2012), viz., 'Alarm' (37 nodes) and 'Insurance' (27 nodes), were not tested. Instead of that, the approach with the proposed threshold was applied to model the recent dataset described in section 4.1. That collected (real) dataset is referred to as $\mathscr{D}_4$.

4.2.2 Experimental setup

The simulated datasets ($\mathscr{D}_{\ell,R}$, $\ell = 1, ..., 3$, $R \in \mathscr{R}$) and the collected one ($\mathscr{D}_4$) were submitted to the processing outlined in Figure 2. For each one of these datasets, $K = 1000$ bootstrap replicas (D_k, $k = 1, ..., 1000$) were modelled as DAGs via Hill Climbing (HC)

[15] The demographic indicators (e.g., 'Age' and 'Sex') are intrinsic to the individual, and they are not dependent on its will. On the other hand, the socioeconomic ones (e.g., 'Occupation' and 'Education') reflect its position in society and are changeable (SCUTARI; DENIS, 2014).

with BDeu metric (the value 10 was the adopted equivalent sample size) - that choice is due to the learner HC, combined to scoring metric BDeu, has shown the best performance in (SCUTARI; NAGARAJAN, 2013). The learners were started from distinct networks, which in turn were obtained from a uniform distribution over *the space of connected graphs* via the algorithm in (IDE; COZMAN, 2002) - the intention here was as avoiding an eventual systematic bias in the learned DAGs well as ensuring a more large space covering.

Given each set of 1000 learned adjacency matrices from each *simulated dataset*, instead of obtaining an overall average matrix ($\bar{E}$ in Figure 2), subsets from 50 to 950 matrices were generated via randomly drawn and, after that, the respective means matrices computed. That strategy resulted in a matrix $\bar{E}$ to each 'artificial' K between 50 and 950 and, in this way, allowed to verify the Equation 3.24 performance more thoroughly. In short, the experiment with simulated data involved to obtain the following expressive amount of adjacency matrices (see E^* in Figure 2):

$$E^* = E^* \left(\ell, R, K, f_{th}\right), \tag{4.2}$$

where $\ell \in \{1, 2, 3\}$, $R \in \mathscr{R}$, $K \in \{50, 51, ..., 950\}$ and $f_{th} \in \{0.01, 0.02, ..., 1.00\}$, and, thus, around 2.7 millions of structures were evaluated.

Regarding the *collected dataset*, a single overall average matrix $\bar{E}$ was computed over the 1000 learned adjacency matrices. The threshold predicted via the proposed formula to $K = 1000$ was 38%. Besides a final adjacency matrix E^* to that cutoff-frequency, more another two ones were gotten, one to a threshold of 35% and another to a cuttof-level of 41%.

4.2.3 Performance evaluation

Following Scutari and Nagarajan (SCUTARI; NAGARAJAN, 2013), the comparisons between the networks learned from simulated datasets and the correct ones (the 'golden models') were performed considering only the underlying undirected graphs. In other words, the focus in the evaluation of the learned networks was on the direct connections and not on the arcs. Since, given a DAG represented by an adjacency matrix E, the transformation

$$U\left(E\right) = E + E^T \tag{4.3}$$

provides the symmetric matrix that represents the underlying undirected graph, to every learned E^* (see Equation 4.2) and its correspondent 'golden' adjacency matrix E^ℓ (remember that ℓ designates the dataset), the following matrices were compared:

$$\begin{aligned} E_u^* &= U\left(E^*\right), \\ E_u^\ell &= U\left(E^\ell\right). \end{aligned} \tag{4.4}$$

Due to symmetry, those comparisons were carried out taking into account only the elements above the main diagonal. Interpreting such elements in every E_u^* as predictions

from a binary classifier and the respective elements in E_u^ℓ as the expected (true) class labels, *confusion matrices* (FAWCETT, 2006; CHICCO, 2017) were built. Once available the components of a confusion matrix, i.e., the *True Positives* (TP), *True Negatives* (TN), *False Positives* (FP) and *False Negatives* (FN), several metrics of performance evaluation could be used. Actually, it is common the simultaneous use of more than one metric derived from such components aiming a more thorough analysis. In (SCUTARI; NAGARAJAN, 2013), for example, the *Accuracy*, *Specificity* and *Sensitivity* (FAWCETT, 2006; CHICCO, 2017) were employed at the same time for assessing the learned networks. An exciting alternative to the simultaneous use of different metrics is the *Matthews Correlation Coefficient* (MCC) (MATTHEWS, 1975), which considers all components of the confusion matrix in its formula

$$MCC = \frac{TP \,.\, TN - FP \,.\, FN}{\sqrt{(TP + FP) \,.\, (TP + FN) \,.\, (TN + FP) \,.\, (TN + FN)}} \tag{4.5}$$

and is robust against eventual imbalanced adjacency matrices (a sparse adjacency matrix, for example) (BOUGHORBEL; JARRAY; EL-ANBARI, 2017). Following the 'tips' in (CHICCO, 2017), in this experiment, the $MCC = MCC\,(\ell, R, K, f_{th})$ was the metric applied to evaluate the networks learned from simulated data.

In Figure 5, the coloured vertical lines depict the obtained confusion matrix to $\ell = 3$, $R = 5$, $K = 950$ along of thresholds f_{th} from 0.0 to 1.0. The correspondent MCC curve is shown in black. Observe that its maximum was higher than 0.9 and occurred around $f_{th} = 38\%$, which is precisely the cutoff-frequency predicted by the formula proposed in Equation 3.24. Note that, according to the definition in Equation 4.5, the highest possible MCC is $+1$ and is reached when the learned network is identical to the 'golden model' (i.e., when both Type-I and Type-II errors are zero).

The curve behaviour in Figure 5 is quite favourable to the proposed formula and noteworthy, but an individual result does not validate it in a broad sense. To comprehensively demonstrate that the Equation 3.24 provides efficient cutoff-frequencies in the learning context outlined in Figure 2, firstly a set $\mathcal{F}_{th}$ of 'good' thresholds to every 'MCC *vs* f_{th}' curve was found as follows:

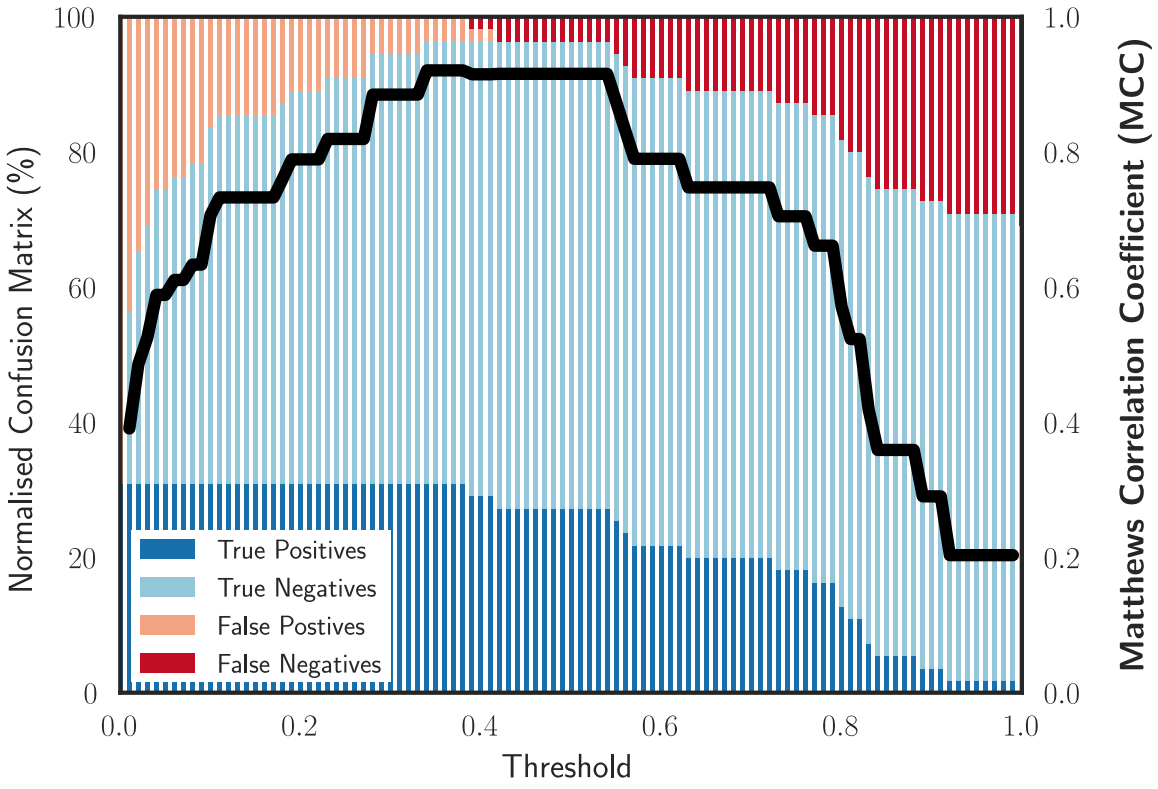
$$\mathcal{F}_{th} = \mathcal{F}_{th}\,(\ell, R, K, \alpha) = \left\{ f_{th} \in [0, 1] \mid (1 - \alpha) \leq \frac{MCC\,(\ell, R, K, f_{th})}{MCC_{max}^{\ell,R,K}} \leq 1 \right\}, \tag{4.6}$$

where (ℓ, R, K) and $MCC_{max}^{\ell,R,K}$ are the *set of parameters* and the *global maximum* of each curve, respectively; the 'α' *relaxation coefficient* establishes the allowed performance loss. After that, over the $\mathcal{F}_{th}$ obtained to each curve, the threshold most close to that one computed via the proposed formula was identified. In other words, the quasi-optimal cutoff-frequency

$$f_{th}^{*}\,(\ell, R, K, \alpha) = \underset{f_{th} \in \mathcal{F}_{th}}{argmin} \left| f_{th} - f_{th}^{+} \right|, \tag{4.7}$$

where f_{th}^{+} is given by Equation 3.24, was obtained to every (ℓ, R, K) presented after the Equation 4.2 and $\alpha \in \{0.01, 0.05, 0.10, 0.15, 0.20, 0.30\}$.

Figura 5: Normalised confusion matrix and respective MCC to $\ell = 3$ ('Sachs'), $R = 5$, $K = 950$ along f_{th} from 0.0 to 1.0.



Source: Author

The collected dataset has not a 'golden structure', and the resulting network from the threshold computed via Equation 3.24 was assessed regarding the captured local associations and the inference capability. Since $K = 1000$, according to the proposed analytical threshold in such an equation, every preponderant directed edge that has occurred with a $f_{\%}^{+} > f_{th_{\%}}^{+} \approx 38\%$ should be accepted as existing (i.e., as sufficiently persistent). For verifying the plausibility of that threshold, the correspondent BN was compared to the ones obtained by adopting 35% ($\approx 1/3 + 1.\sigma_{f^+}$) and 41% ($\approx 1/3 + 5.\sigma_{f^+}$). In these comparisons, each BN was evaluated regarding both the number of directed edges (structural complexity) and also the distribution of success rates in the predictions (inferences) of the *Cognitive Impairment*[16] variable (see the description of the node CI in subsection 4.1.1.4). In addition to the three DAGs reduced from the ensemble of structures (that were learned from the bootstrap replicas), a tree structure was learned via Chow-Liu

[16] From the medical perspective, it is fundamental to infer the CI node from others measures - the *Cognitive Impairment* is an essential predictor of severe dementias such as the Alzheimer's disease (ALBERT et al., 2011). Thus, the BNs learned from the collected dataset using three different thresholds (35%, 38% and 41%) were compared regarding the performance in inferring the CI node.

algorithm (CHOW; LIU, 1968). The correspondent 'Chow-Liu BN'[17] was used merely to verify the consistency of the inferences to the node CI performed with the three BNs mentioned above.

In order to build the four BN models to the collected dataset (the four structures were learned before and with all the data) and at the same time to assess their performances in prediction of CI node, the data in $\mathscr{D}_4$ were randomly partitioned into two complementary subsets: 75% of the collected data (150 rows) for training the parameters (i.e., for learning the conditional probability table - CPT) and the others 25% (50 rows) for obtaining the success rate in predictions of the node CI (i.e., the number of correct predictions in the 50 tests). That one round of cross-validation was repeated 1000 times for associating a distribution of success rates to each model.

4.2.4 Resources

For all computational tasks, a computer with an i7 7th Gen Processor, 16 GB of RAM and Linux Based Operating System was used. The algorithms were implemented in Python language with the aid of scientific computation libraries such as Numpy, Scipy, and Matplotlib (OLIPHANT, 2007). For general networks manipulation, the NetworkX (HAGBERG; SCHULT; SWART, 2008) was employed. The structures to the collected data and those to the simulated ones were learned via PEBL (SHAH; WOOLF, 2009) and PGMPY (ANKAN; PANDA, 2015) libraries, respectively. For both the learning of the networks' parameters (CPTs) and the statistical inferring of the CI node, the Pomegranate package (SCHREIBER, 2018) was applied. Particularly, the Chow-Liu structure (a tree) was learned via Pomegranate.

The computational time to the learning of structures to the simulated data was around 700 hours (almost a month), and the worst case has occurred to the 'Sachs' - the time to learn only the networks to this dataset has corresponded to 75% of the total. In this point, it is crucial to mention that no restriction (e.g., the number of parents to each node) was imposed to reduce the vast space (see Figure 1) to be searched by the greedy algorithm.

[17] Despite it is an inappropriate reference in the capture of associations between variables (it always is a tree), the Chow-Liu BN can be rapidly trained (structure and parameters) and is efficient in predictions tasks (fast and accurate).

5 RESULTS AND DISCUSSION

This chapter presents the results of each experimental part (with the correspondent discussion) in an own section. First, the results (and its analysis) related to the medical application are presented (see section 4.1), and, after that, the ones associated with the validation of the analytical threshold (see section 4.2).

5.1 Part I: Results and discussion to the real application of the adapted approach

In Table 1, the baseline characteristics of the 200 elderly participants are presented. The two first moments (i.e., $\hat{\mu}$ and $\hat{\sigma}$) were shown only to provide a rough approximation of the measurement distributions. However, the data were not normally distributed[1] according to the Shapiro-Wilk Test[2] performed on the collected data. Only the weights were drawn from a Gaussian distribution (a 5% confidence level was used for the tests of normality).

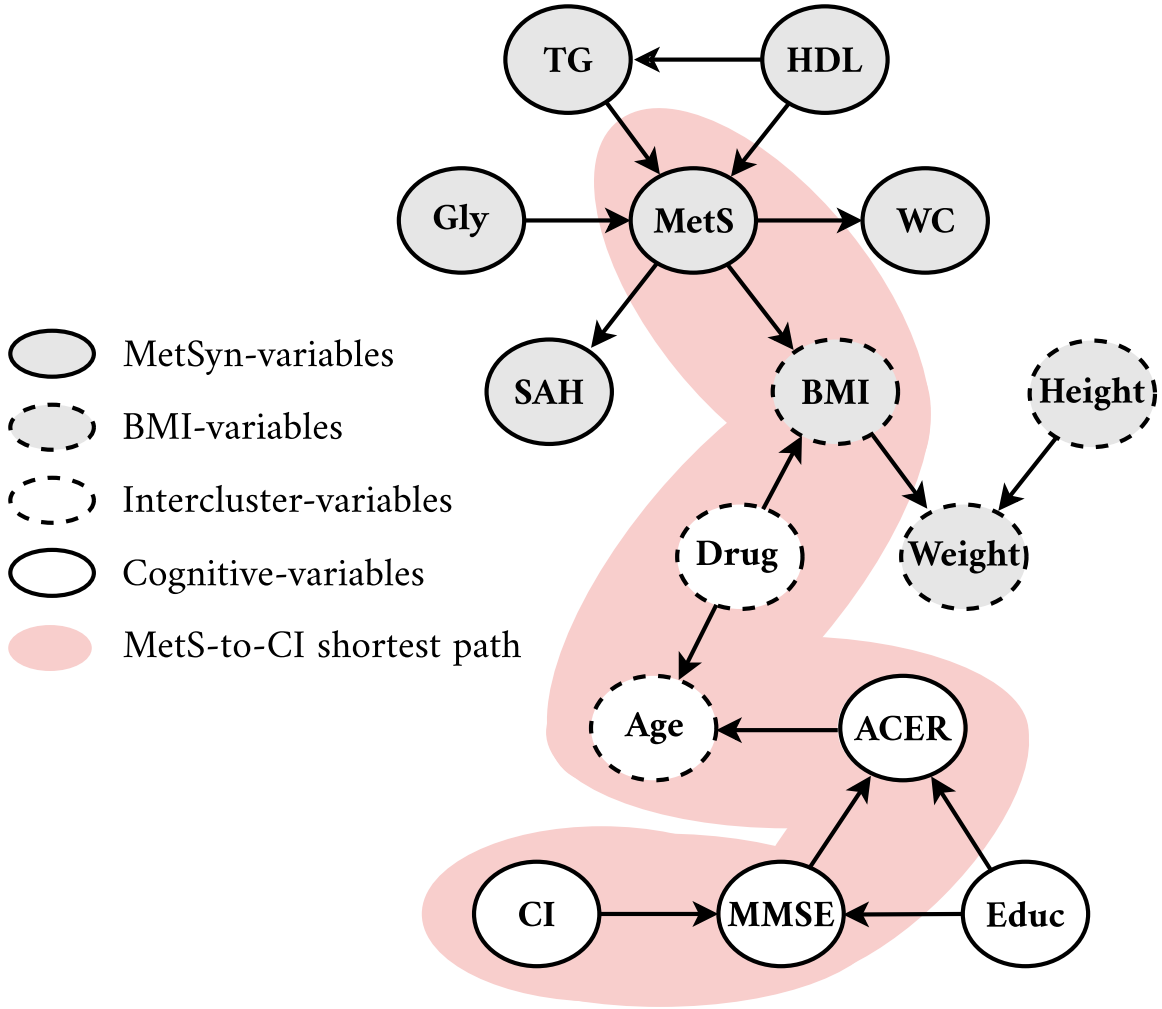
Tabela 1: Overall statistics of the collected dataset. The second column shows the short names of correspondent nodes.

Feature	**Node**	**Overall**
Age (yr)	Age	72 ± 6
Sex ($men : women$)	Gender	$63 : 137$
Formal Education (yr)	Educ	5 ± 4
Physical Activity (%)	Exercise	63 (31.5%)
Pharmaceutical Drug Use (%)	Drug	174 (87.0%)
Smoking (%)	Smoking	72 (36.0%)
Alcoholic Intake (%)	AlcBev	24 (12.0%)
Weight (kg)	Weight	69.4 ± 13.4
Height (m)	Height	1.60 ± 0.07
Body Mass Index (kg/m^2)	BMI	27.14 ± 4.89
Waist Circumference (cm)	WC	91.1 ± 13.1
Triglycerides (mg/dL)	TG	128.2 ± 56.8
Glycemia (mg/dL)	Gly	101.6 ± 28.8
High-Density Lipoproteins (mg/dL)	HDL	76.4 ± 35.6
Systemic Arterial Hypertension (%)	SAH	132 (66.0%)
IDF-based Metabolic Syndrome (%)	MetS	79 (39.5%)
Addenbrooke's Cognitive Examination-R ($score$)	ACER	66 ± 16
Mini-Mental Status Examination ($score$)	MMSE	23 ± 4
MMSE-based Cognitive Impairment (%)	CI	147 (73.5%)

In Figure 6, the BN structure model learned from the gathered dataset is shown. This graph does not contain all the nodes specified in Table 1 because the ones that

[1] The obtained BN was discrete and the employed structural learning was fully non-parametric.
[2] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html

Figura 6: BN structure learned from the collected dataset using the PEBL computational package. The shaded region highlights the existence of an indirect link between MetS and CI. Along the path from the MetSyn-variables (MetS and the ones that define it) to the Cognitive-variables (CI and the ones that define it), only two basic connections are of the converging type: they occur in the BMI and Age nodes.



Source: Author

remained isolated from the two focus variables (MetS and CI) were excluded. The shaded region highlights the main finding: the existence of a connection path from MetS to the cognitive nodes (CI, MMSE-score, and ACER-score).

The BN structure (DAG) in Figure 6 is consistent with the definitions of the variables (nodes) in the subsection 4.1.1. The determinant nodes appear exactly around the determined node (i.e. a well-defined function of the other nodes) in such a way that it

is possible to see clusters like one of the MetSyn-nodes (MetS and the nodes that define it) and one of the Cognitive-nodes (CI and the nodes that define it). The BMI node and its determinant components ('Weight' and 'Height' nodes) also appear in the learned DAG as a (small) cluster.

The cluster BMI→Weight←Height is a basic converging connection where the common child (or 'collider') is the node 'Weight', and its parents are the BMI and 'Height' nodes. These two nodes are then D-separated by that one. In the dataset as a whole, there is no significant direct dependence between BMI and 'Height', but it arises when the collected data is segregated in terms of 'Weight' values (as explained in the section 2.3, the instantiation of the intermediate node is the unblocking condition for a converging connection). In fact, an eventual absence of dependence between the variables BMI and 'Height' for a given 'Weight' (conditional dependence) would be an inconsistency since the curve of BMI *versus* Height, in its definition[3], is parameterized by the 'Weight'. The lack of a direct arc between BMI and 'Height' (i.e. the absence of an unconditional dependence between these two variables) reinforces the credibility of the learned DAG since such a lack agrees with the following assumption commonly adopted when using BMI as an obesity indicator (KEYS et al., 2014; DIVERSE, 2005): BMI is strongly correlated with 'Weight', but is independent of 'Height'. In general, $Weight \propto Height^{\alpha}$, with $\alpha \approx 2$ (EKNOYAN, 2008), therefore the direct dependence between BMI and 'Height' in a population is weak at the most[4].

The indirect path highlighted in Figure 6 implies the association MetS-CI is conditioned to the instantiation of BMI (or 'Weight') and 'Age'. In practice, this means that it is necessary to segregate (filter) the dataset in terms of these two 'colliders' in order to observe a dependence between MetS and CI (see item 'converging connection' in the section 2.3). Since the CI, MMSE, and ACER nodes are well-clustered in the obtained DAG, these considerations of the MetS-CI association are also valid concerning the MetS-MMSE and MetS-ACER associations. Note that the absence of a v-structure between such Cognitive-variables and 'Age' expresses the well-known absolute dependence between cognition and aging (WHO, 2015).

The need for stratification coded in the path MetS-CI by the two 'colliders' (BMI and 'Age') is not a mere statistical oddity. Such a finding is substantially relevant since it informs that, in elderly individuals, the role of metabolic syndrome components in cognitive

[3] The well-known equation BMI=Weight/Height^2.

[4] Adolphe Quetelet (1796-1874) performed early investigations concerning the relationship between 'Weight' and 'Height' in a population and obtained a value for the slope of the regression line $log(Height)$ x $log(Weight)$ close to 2 (EKNOYAN, 2008). More recent studies (e.g. in (DIVERSE, 2005)) identified populations with an exponent α slightly different from 2 - for those populations, in fact, some dependence can be expected between 'Height' and BMI. However, in such cases, this index may not be reasonable for comparisons of obesity between individuals with different heights.

performance changes with different age and obesity segments (BMI is an obesity index). This has profound implications for practice. For example, to mitigate cognitive decline, this finding supports that a nutritional recommendation for younger elderly individuals may not be applicable for older ones, and vice-versa. From the preventive policies viewpoint, such a result suggests interventions in the elderly to mitigate the cognitive decline should be specific to each obesity (BMI) range and each age segment.

The data segregation induced by the 'colliders', BMI and 'Age', can also explain some results seen in the literature regarding MetS-CI linking:

- the investigation in (ARAÚJO et al., 2017) probably did not find a dependence between the two focus variables because the assessment was performed in the dataset as a whole, i.e. without any segregation in terms of 'Age' and BMI (or 'Weight') as suggested by the MetS-CI path in Figure 6;

- the study in (LIU et al., 2013) concluded that MetS influences the cognitive performance according to the age segment and even found a better cognitive score for older elderly individuals ($\approx$ 80 years) with metabolic syndrome. The network in Figure 6 is consistent with this, since the 'collider' in 'Age' means that the dependence between MetS and CI is conditional on the age stratum. On the other hand, the same study did not infer BMI stratification as essential to the relationship between MetS and CI - the structure in Figure 6 disagrees with this since BMI is also a 'collider' between MetS and CI. In (LIU et al., 2013), both the mean and variance of the BMI are much smaller than those in this study (23.8 ± 3.2 compared to 27.14 ± 4.89). That high homogeneity can be interpreted as an implicit instantiation of the BMI which in turn would explain why such a study found a dependence between MetS and CI without any explicit conditioning of the BMI. Thus, given such an interpretation, the inconsistency above disappears.

The study in (DEARBORN et al., 2014) disagrees with the result in Figure 6 as it found an overall association between MetS and CI. Such opposition is not necessarily a controversy in itself because the population sample in such a study is complementary to that used here concerning the age range (45 to 64 years against 65 to 90 years). The research in (FENG et al., 2013) concluded that MetS and its principal component - central obesity - are associated with CI. Since BMI is an obesity index, the result of such research can mean that the MetS-CI linking is dependent on the BMI instantiation, similar to that indicated in Figure 6. The work in (TOURNOY et al., 2010) agrees with this study regarding the absence of overall dependence (without stratification) between MetS and CI. However, a more detailed comparison is not viable since such work uses a sample that primarily contains individuals much younger than the subjects enrolled in this study (40 to 79 compared to 65 to 90).

The investigation in (LAUDISIO et al., 2008) highlighted a dependence between MetS and CI dependent on the stratification of the ages. Although such a finding is consistent with the BN structure in Figure 6, the same investigation did not find a significant relationship between MetS and CI for male individuals. Clearly, the influence of sex on MetS-CI linking was not captured by the DAG in Figure 6. A possible explanation for such a discrepancy could be the difference between the cognitive assessment protocols, since (LAUDISIO et al., 2008) utilized the Hodkinson Abbreviated Mental Test (AMT) instead of the MMSE. However, in a general way, such protocols capture practically the same features related to cognitive decline (LAUDISIO et al., 2008), and perhaps a better reason for such discordance may be the conservative strategy itself that was adopted here for learning the network in Figure 6 (see below).

A limitation in this application is the probable exclusion of essential nodes by the high confidence level adopted (67% threshold). The above effect of sex on the MetS-CI relationship, for example, perhaps may have been captured by adopting a lower threshold[5]. The 67% level was a 'worst case' strategy used as a simple way to avoid any doubt as to the real existence of the connections and primarily ensure the credibility of the path from MetS to CI seen in the result shown in Figure 6. The use of an increased threshold is also due to the possible social impact of this important finding; without a large safety margin, it would not be correct to state a dependence that is possibly capable of influencing prevention policies for dementias.

To get a more comprehensive model, capable of being utilized for medical prediction purposes in new patients[6], a potentially better strategy to learn the structure would be to adopt a threshold closer to 33% for each edge type ('$\leftarrow$', '$\rightarrow$', and 'absent'). After this medical application, efforts in that direction resulted in the closed formula proposed in the Equation 3.24, which in turn provided a cutoff-frequency that has indeed resulted in a network with high performance regarding prediction of the node CI (see Figures 9 and 11).

An undesired aspect in the present application is the relatively small sample size. A model-averaging with data perturbation approach was applied in the structural learning to partially compensate for such scarcity. At first glance, the convenience sampling also seems a critical aspect of this application since that such methods of data gathering can eventually induce some bias and compromise the generalization of results (JAGER; PUTNICK; BORNSTEIN, 2017). However, the convenience sampling was performed across multiple Health Units that were spread uniformly throughout the city, giving a wide range of profiles. In this way, the collection procedure as a whole is not so far from an ideal

[5] Posterior tests with the threshold predicted by Equation 3.24 has revealed that gender indeed affects the link between MetS and CI. See Figure 9.

[6] In this context, medical prediction means to automatically provide the possible state of a medical interest node of a new patient through measurements of only some other nodes of this same new patient. See Figure 11.

random sampling, and it seems plausible to assume that the bias due to sampling is in fact sufficiently weak[7].

5.2 Part II: Results and discussion to the validation of the proposed threshold

In Figure 7, the quasi-optimum thresholds appeared around the black curve computed via Equation 3.24, even when a performance quite close to the maximum was demanded ($\alpha = 1\%$) - such behaviour suggests that on average the proposed threshold is a suitable choice. As can see, the variability regarding the proposed curve has been dependent on the data and was greater to the 'survey' and 'asia' than to the 'sachs' (bypassing the tiny bubbles). On the other hand, such dependence has reduced rapidly as the tolerance was increased - for an 'α' of only 10%, most of the quasi-optimum thresholds already has adhered to the suggested curve. Although the rigorously optimum threshold has shown dependence on the data and such fact is recognised in the literature (FRIEDMAN; GOLDSZMIDT; WYNER, 1999; SCUTARI; NAGARAJAN, 2013; BROOM; DO; SUBRAMANIAN, 2012), the fast convergence to the analytical curve (i.e., the overall collapse on the proposed curve when the 'α' still can be considered as insignificant) suggests that, in general, the adoption of the proposed threshold implies in a marginal performance loss.

Still regarding the Figure 7, the quasi-optimal thresholds for $R << 1$ (scarce data) and $R >> 1$ (massive data) were represented by tiny and huge bubbles, respectively. Speaking about the data availability, when the amount of modelled instances has been around or higher than the number of unknown free parameters ($R \geq 1$), the quasi-optimal thresholds tended to the proposed curve as the performance exigence was relaxed via 'α'. On the other hand, when few instances were used in the presence of many unknown parameters ($R < 1$), even to the maximum admitted tolerance (30%), a large number of quasi-optimal thresholds remained out of the proposed curve.

Although the results in Figure 7 indicate that in general the Equation 3.24 provides cutoff-frequencies able to recover the most edges among nodes correctly, the adopted approach (see Figure 2) does not ensure a complete absence of cycles in the learned networks. Some preliminary tests have revealed acyclicity already at low frequencies (i.e., too below the suggested curve) and, since it is impossible to generate new cycles by removing arcs, the existence of them was initially assumed as quite improbable around

[7] In this context, it is also worth mentioning that one of the principal purposes of the Health Units used as sampling sites is to provide preventive assistance. In Brazil, the focus in such basic Health Units is not emergency or solely corrective assistance. Thus, the elderly do not necessarily go to such basic Health Units only when they are affected by suspect symptoms. That aspect significantly mitigates the possibility of selection bias related to the patient group. According to (COOPER, 1995), mitigation is essential since such bias could result in a network with 'spurious arcs' (i.e. direct associations that would not be induced if the entire elderly population were considered). See the interesting example on page 143 of (COOPER, 1995).

the cutoff-frequencies close to proposed curve. However, posteriorly, counting the number of cycles according to the increase of the cutoff-frequency to every possible combination of (ℓ, R, K), a somewhat different result of the expected one was gotten and is shown in Figure 8. As can be seen in Figure 8, many of adjacency matrices learned from the 'Sachs' dataset became acyclic only to cutoff-frequencies above the analytical curve. In other words, considering the 'Sachs', in many occasions the use of the proposed threshold has resulted in cyclic networks, notably to $R > 1$ (more data than free parameters) and $K > 400$ (massive amount of bootstrap replicas). Curiously, the 'Sachs networks' have been justly the closest to the 'golden model' when the proposed formula was applied (see the Figures 5 and 7 again).

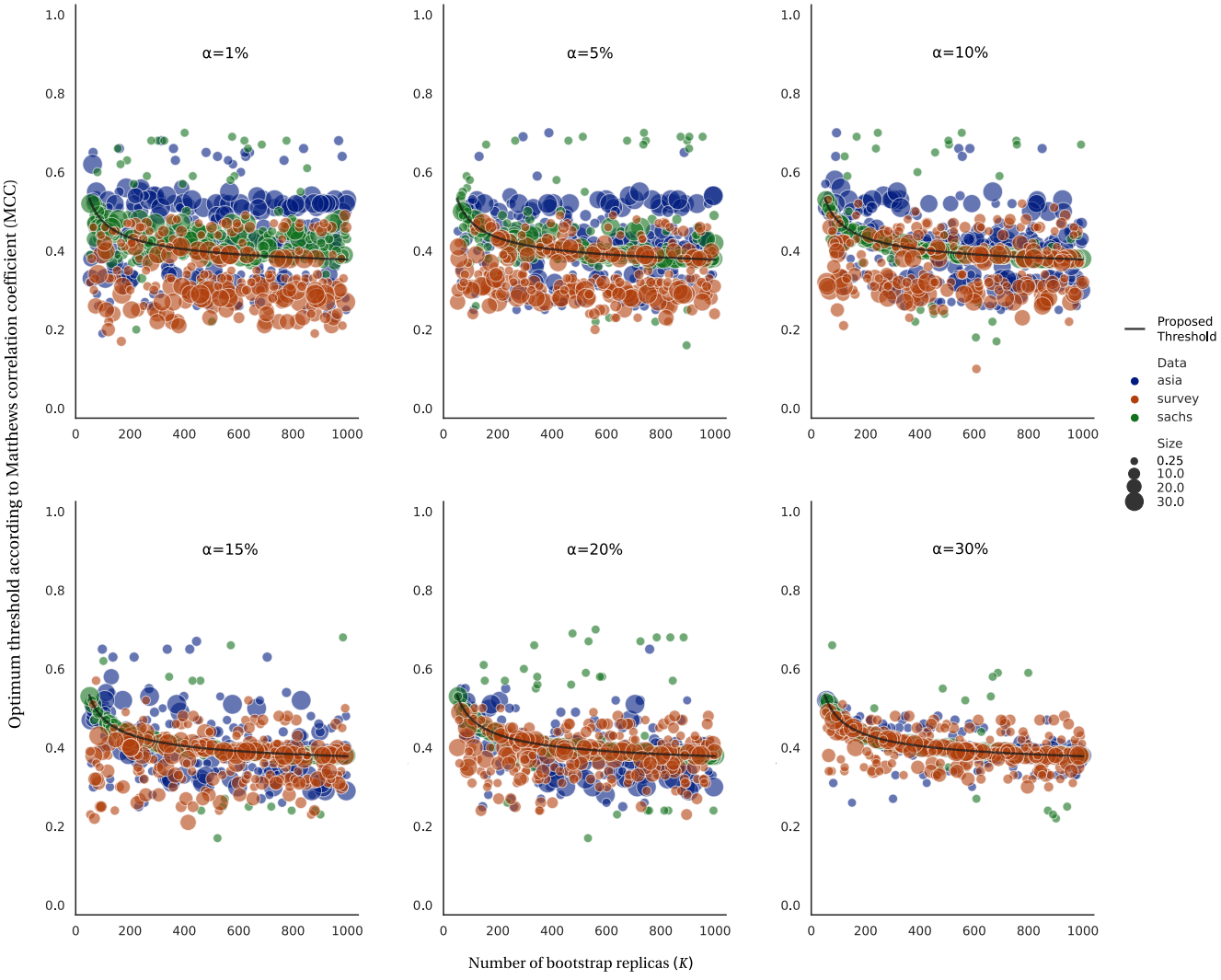
In order to preserve the performance (regarding edges and not arcs) achieved with the analytical threshold and eliminate possible oriented loops due to an eventual insufficient cutoff-frequency to generate a DAG, a viable strategy is to invert just an arc of each found cycle. Since in the adopted approach the chance of a wrong choice to an edge orientation increases as the frequencies of its two arcs close, it seems reasonable to invert that one connection whose opposed arcs present the closest occurrence frequencies. In short, eventual directed cyclic graphs resulting from the thresholds computed with the proposed formula are entirely manageable.

The results in Figures 9 to 12 to collected data are also exciting since the local relationships seem consistent (Figure 9) and the cost-effective has been better to the proposed threshold than to its neighbourhood (Figures 11 and 12). The directed graph in Figure 9 does not have cycles and is locally consistent in the sense that determinant and determined nodes are neighbours in the structure - the detailed definition to every variable (node) and the functional relations can be verified in the section 4.1 (in particular, see subsections 4.1.1.3 and 4.1.1.4).

In Figure 11, one can see that the reduction of the ensemble via the proposed threshold of 38% resulted in a BN (Figure 9) that achieved better performance than the BN of Chow-Liu (Figure 10) regarding success rates in the prediction of the node CI: the pair $(\hat{\mu}_{CI}, \hat{\sigma}_{CI})$ was $(0.8128, 0.0507)$ for BN of Chow-Liu against $(0.8452, 0.04890)$ for BN reduced from the ensemble by using the threshold of 38%. Regarding the performance of different thresholds, it got worse for 41% and remained practically the same for 35%: the pair mentioned above was $(0.8118, 0.0511)$ for the higher threshold (41%) and it was $(0.8500, 0.0508)$ for the lower one (35%). The boxplot for 35% seemed (see each boxplot in Figure 11) quite similar to the one for 38%, and, then, the Kolmogorov-Smirnov statistic[8] was computed based on the two correspondent samples (i.e., the well-known nonparametric K-S test was performed). The resulting values were $KS_{statistic} = 0.0310$ and $p_{value} = 0.7161$, and, in fact, the null hypothesis that the two distributions has been

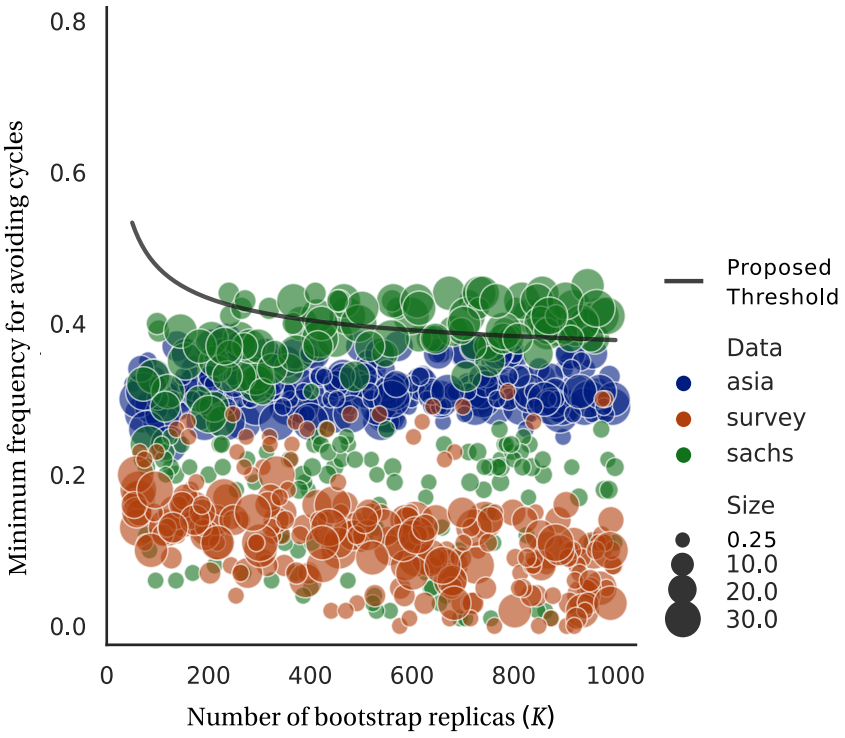
[8] The $KS_{statistic}$ was computed via the "$scipy.stats.ks\_2samp$" function.

Figura 7: Coalescence trend of the 'good' cutoff-frequencies on the analytical curve proposed to predict them. The thresholds surrounding such a curve adhered to it soon that the tolerance 'α' was slightly increased, except to the cases in which $R < 1$, i.e., when the number of instances has been smaller than the number of free parameters in the 'golden model'. In fact, many of the tiny bubbles are observed at most nearby from the analytical curve, and not on it, even to a relaxation of 30%.



Source: Author

Figura 8: Minimum cutoff-frequency for avoiding cycles to each combination (ℓ, R, K). The green bubbles above the analytical curve mean that in such cases the use of the proposed threshold would result in directed networks with cycles (i.e., they would be not DAGs). Since a cycle can be eliminated by inverting just one arc, its eventual presence is manageable.

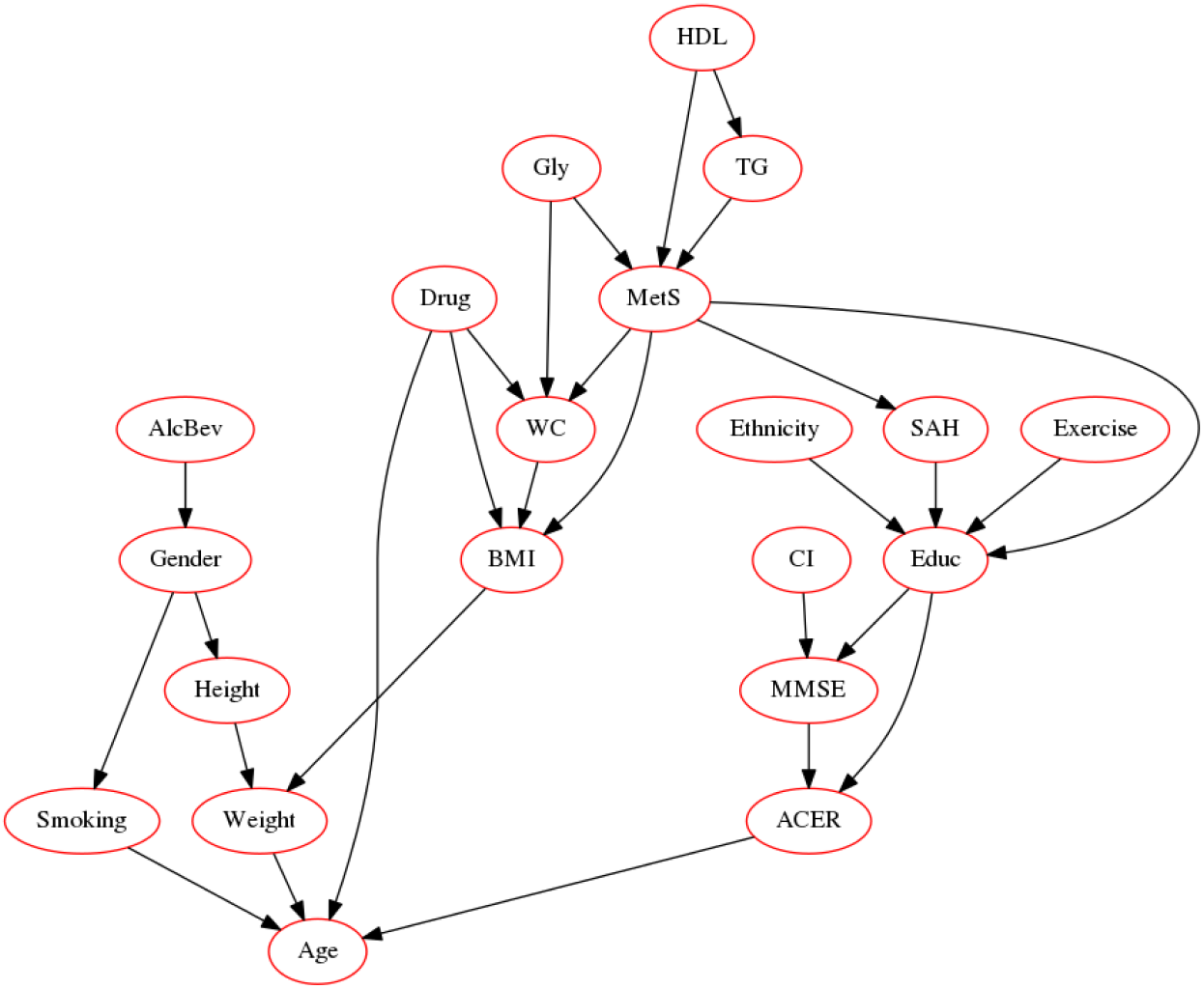


Source: Author

the same cannot be rejected.

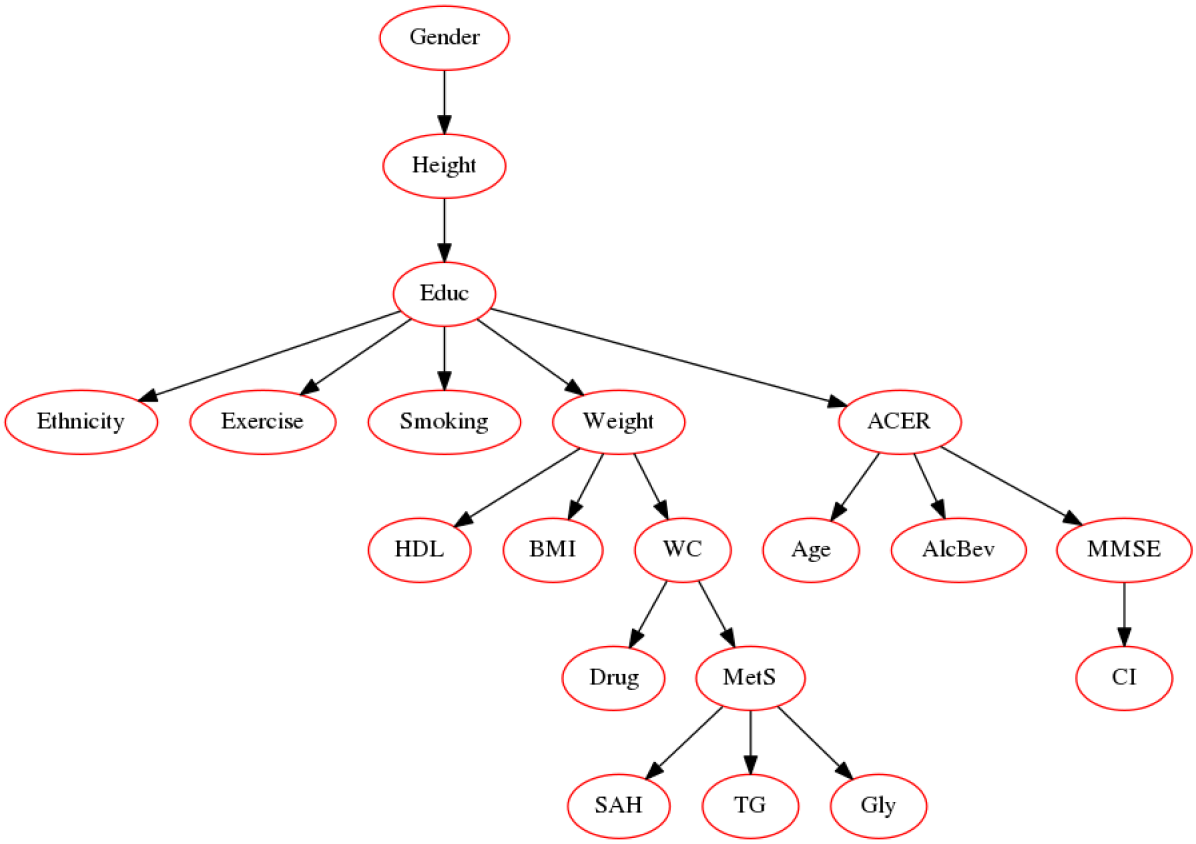
In Figures 11 e 12, one can see that there is a kind of trade-off between prediction performance and structural complexity as the threshold variates. From a simultaneous analysis of these two figures, one can see that the threshold of 38% obtained via Equation 3.24 is a quite reasonable choice: if the threshold decreases from 38% to 35%, the gain in terms of success rates will be negligible (in fact, the K-S test mentioned above did not distinguish the two correspondent distributions) at the cost of an increase in complexity of at least 14% (from 28 to 32 arcs); and, if the threshold increases from 38% to 41%, the decrease in complexity will be only of two arcs at the cost of lowering the BN's prediction performance to a level already achieved by Chow-Liu's BN itself.

Figura 9: BN structure obtained from the ensemble by using the proposed threshold of 38%.



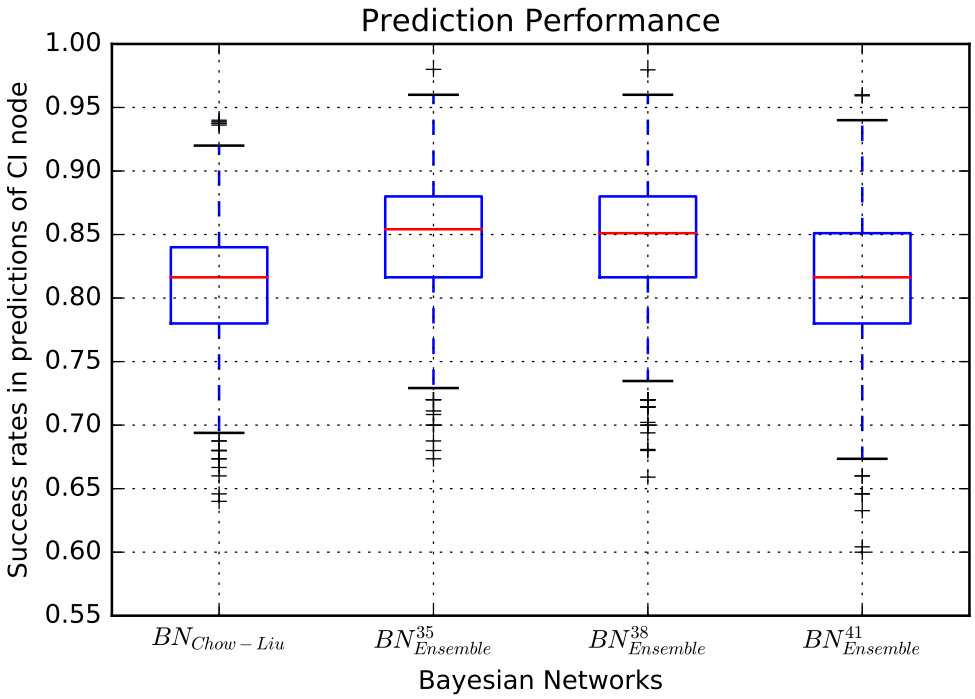
Source: Author

Figura 10: BN structure provided by the Chow-Liu algorithm. Observe the use of a severe restriction to reduce the search space: each node has not more than a parent.



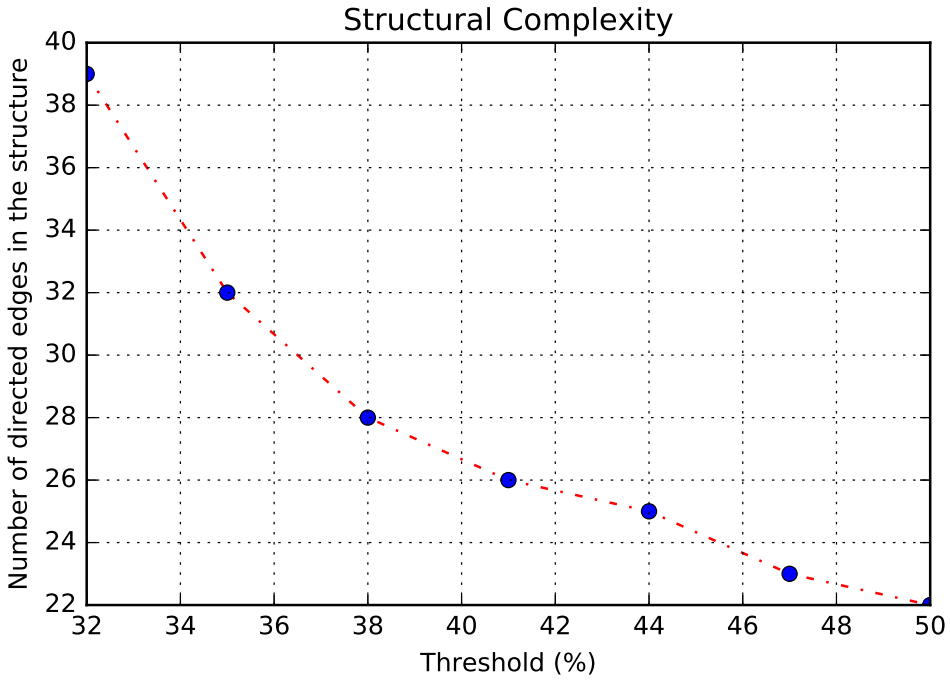
Source: Author

Figura 11: Boxplot of the success rates for node CI according to BN model used in the predictions.



Source: Author

Figura 12: Structural complexity regarding the number of arcs for thresholds around 38%.



Source: Author

6 DISSEMINATION ACTIVITIES

The following papers resulted from this doctoral research:

- **Journals**:
  - **Published**: Tadeu Jr Gross, Renata B. Araújo, Francisco A. C. Vale, Michel Bessani, Carlos D. Maciel, Dependence between Cognitive Impairment and Metabolic Syndrome Applied to a Brazilian Elderly Dataset, **Artificial Intelligence in Medicine**, Elsevier, v. 90, p. 53-60, 2018.
  - **Submitted**: Tadeu Jr Gross, Michel Bessani, Willian Darwin Jr, Renata B. Araújo, Francisco A. C. Vale, Carlos D. Maciel, An Analytical Threshold for Combining Bayesian Networks, **Knowledge-Based Systems**, Elsevier, 2018.
- **Conference**:
  - **Published**: Tadeu Jr Gross, Michel Bessani, Jonas R. Dourado, Daniel R. Lima, Carlos D. Maciel, Evaluation of Long-Term Feature Parameters for Noise-Tolerant Speaker Identification. **Proceedings XIII SBAI**, 2017.

The below papers are not related to this Thesis and were just collaborations:

- **Journals**:
  - **Submitted**: Jonas R. Dourado, Michel Bessani, José R. Monteiro, Tadeu Jr Gross, Willian Darwin Jr, Carlos D. Maciel, Parallelism Strategies for Big Data Delayed Transfer Entropy, **Entropy**, 2018.
  - **Submitted**: Luiz D. Neto, Michel Bessani, Rodrigo Z. Fanucchi, Tadeu Jr Gross, Carlos D. Maciel, A Multi-objective Swarm Intelligence Approach for Field Crews Patrol Optimisation in Power Distribution Systems Restoration, **Journal of Control, Automation and Electrical Systems**, Springer, 2018.
- **Conference**:
  - **Published**: Luiz D. Neto, Michel Bessani, Rodrigo Z. Fanucchi, Tadeu Jr Gross, Carlos D. Maciel, Particle Swarm Optimisation Applied to the Routing of Electrical Maintenance Teams. **Proceedings XIII SBAI**, 2017.

7 CONCLUSION

In this doctoral research, via the proposed learning strategy, the existence of conditional dependence between metabolic syndrome (MetS) and cognitive variables (CI, MMSE-score, and ACER-score) in an elderly sample was identified - a significant medical finding. The learned structure suggests that in elderly individuals the role of metabolic syndrome (and its components) in cognitive performance can non-trivially change over different BMI (or 'Weight') and age ranges. In a subsequent study, investigation of the collected dataset and the specific nature of the MetS-CI relationship for each combination between segments of age and BMI is planned.

In general, it was noted that a BN structure allows for researchers to obtain valuable insights concerning the properties of a dataset using visual inspections on its graphical representation (DAG) - the D-separation concept plays a central role in that knowledge extraction process. In this context, it is also worth highlighting that a data-driven structural model (DAG) can be used as a reliable 'guide' to define which variables should be considered in an eventual coupling equation (linear or nonlinear) for predictions.

A BN structure captures both linear and nonlinear relationships and, in comparison to other statistical approaches that can eventually be insensitive to nonlinear couplings, it can provide different and more robust implications. Due to that broad capability, it would be quite promising to directly compare the BNs (structures) from several existing similar medical datasets around the globe.

A still more powerful strategy would be to label each existing dataset with the respective region and then merge them for the training of a single BN model. The global structure would enable a systematic identification (via D-separation) of regional attributes eventually capable of changing the associations nature between cognitive decline and known risk factors. Note that the findings already reported in the literature for the isolated datasets would be obtained directly from such a global model via statistical marginalization.

In this doctoral research, the adopted averaging strategy is a version modified from the traditional one, and the significant contribution is that it allows the application of an analytical threshold instead of '*ad hoc*' or numerical ones[1]. Traditionally, the merging of structures learned from bootstrap replicas is performed taking into account the edges persistence. In the adopted alternative approach, the arcs stability has been the feature observed along the networks ensemble.

[1] In the capture of the MetS-CI link, despite the employ of the approach proposed in section 3.1, an '*ad hoc*' (i.e., an increased) cutoff-frequency was used to practically eliminate the chance of a false positive in an as serious issue.

A critical aspect (and perhaps the most important) in the reduction of a networks ensemble to a single robust topology is to set the minimum frequency of occurrence that legitimates the considered feature. In this way, the principal concern in this study has been to establish an appropriate threshold to accept the most recurrent arc between each pair of nodes as authentic. The efforts to this end resulted in a closed-form expression to compute that threshold. The curve to that formula, viz. Equation 3.24, is presented in black in Figure 7 and one can see that it is narrowly surrounded by the 'good thresholds'.

That analytical cutoff-frequency to eliminate spurious arcs was derived via an analogy to a three states random-walk. The effectiveness of the proposed formula (and also of the adopted learning approach itself) was verified on three simulated datasets and also on a recently collected (real) dataset. In the case of simulated data, the 'golden models' were available, and the predicted cutoff-frequencies via the proposed formula were compared to proper thresholds considering different numbers both of instances and bootstrap replicas. Regarding the collected dataset, the resulting BN was assessed concerning both inference capability and local relationships recovering.

The selected benchmark datasets have 11 nodes at most because, besides the limited computational resources (see subsection 4.2.4), in every greedy search, no restriction was imposed to reduce the enormous space of candidate networks (see Figure 1) - the computational learning of all networks related to simulated data took almost a month ($\approx$ 700 hours). Benchmark datasets with more nodes still will be tested via parallelisation since the learning of structures from bootstrap replicas is an embarrassingly parallel task (see Figure 2.a).

The exciting result to the simulated data was the coalescence of the 'good thresholds' on curve predicted by the proposed formula (see Figure 7). That adherence phenomenon was observed along of all ensemble sizes (i.e., all different amounts of modelled bootstrap replicas) and it has been coarse only when the data have been extremely scarce (i.e., when the number of used instances have been fewer than the number of free parameters). Regarding the proposed approach (and its analytical threshold) applied to the collected dataset, the resulting BN (see Figure 9) has encoded in its structure the expected local relationships among the nodes and at the same time has been able to perform predictions of a relevant node with high success rates (see Figure 11).

Noisy data are ubiquitous in the real world and to detach the measurable 'signal' from the 'noise' is essential in the knowledge acquisition from the raw data. That means that often an 'isolated' network model has as genuine dependencies well as spurious ones. The learning of multiple structures from bootstrap replicas and the posterior reduction of the ensemble to a single topology is a known effective method to learn a robust network (i.e., that encodes correctly the relevant information that is latent in the data).

The discarding of spurious dependencies via networks averaging is an attractive

strategy which so far has been applied only using '*ad hoc*' or numerical cutoff-frequencies. In this aspect, the conducted research contributes with a formula to compute efficiently and fastly a proper threshold when the average is performed based on arcs persistence. Furthermore, since the averaging technic often applied in literature considers the edges strength, the possibility of the validated approach (associated with the analytical threshold) to result in more proper structure models in some situations cannot be discarded, and more comprehensive comparisons still need.

REFERÊNCIAS

ALBERT, M. S. et al. The diagnosis of mild cognitive impairment due to alzheimer's disease: Recommendations. **Alzheimer's & Dementia**, Elsevier, v. 7, n. 3, p. 270–279, 2011.

ALBERTI, K. G. M.; ZIMMET, P.; SHAW, J. The metabolic syndrome: a new worldwide definition. **The Lancet**, Elsevier, v. 366, p. 1059–1062, 2005.

ALEXIOU, A. et al. A bayesian model for the prediction and early diagnosis of alzheimer's disease. **Frontiers in Aging Neuroscience**, v. 9, p. 77, 2017.

ALONSO-BARBA, J. I. et al. Scaling up the greedy equivalence search algorithm by constraining the search space of equivalence classes. **International Journal of Approximate Reasoning**, v. 54, n. 4, p. 429 – 451, 2013.

ANKAN, A.; PANDA, A. **Mastering Probabilistic Graphical Models Using Python**. [S.l.]: Packt Publishing Ltd, 2015.

ARAÚJO, R. B. et al. Memory complaint, cognitive disorder, chronic disease and life habits in adults and elderly of a brazilian urban community. **Alzheimer's & Dementia: The Journal of the Alzheimer's Association**, Elsevier, v. 12, p. P1070, 2016.

_____. Relationship between cognitive disorder and metabolic syndrome in elders in a brazilian community. **Alzheimer's & Dementia: The Journal of the Alzheimer's Association**, Elsevier, v. 13, p. P721, 2017.

AUSSEM, A.; MORAIS, S. R. de; CORBEX, M. Analysis of nasopharyngeal carcinoma risk factors with bayesian networks. **Artificial Intelligence in Medicine**, v. 54, n. 1, p. 53–62, 2012.

BARTLETT, M.; CUSSENS, J. Integer linear programming for the bayesian network structure learning problem. **Artificial Intelligence**, v. 244, p. 258 – 271, 2017.

BERG, E. Van den et al. The metabolic syndrome is associated with decelerated cognitive decline in the oldest old. **Neurology**, AAN Enterprises, v. 69, p. 979–985, 2007.

BESSANI, M. et al. Evaluation of a dental caries clinical decision support system. In: **Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC)**. [S.l.: s.n.], 2017. p. 198–204.

BIELZA, C.; NAGA, P. L. Bayesian networks in neuroscience: a survey. **Frontiers in Computational Neuroscience**, v. 8, p. 131, 2014.

BORNSTEIN, M. H.; JAGER, J.; PUTNICK, D. L. Sampling in developmental science: Situations, shortcomings, solutions, and standards. **Developmental Review**, Elsevier, v. 33, n. 4, p. 357–370, 2013.

BOUGHORBEL, S.; JARRAY, F.; EL-ANBARI, M. Optimal classifier for imbalanced data using matthews correlation coefficient metric. **PLOS ONE**, Public Library of Science, v. 12, n. 6, p. 1–17, 06 2017.

BROOM, B. M.; DO, K.-A.; SUBRAMANIAN, D. Model averaging strategies for structure learning in bayesian networks with limited data. **BMC Bioinformatics**, v. 13, n. 13, p. S10, Aug 2012.

BRUCKI, S. et al. Suggestions for utilization of the mini-mental state examination in brazil. **Archives of Neuropsychiatry**, SciELO Brazil, v. 61, p. 777–781, 2003.

CAMPOS, C. P. de et al. Entropy-based pruning for learning bayesian networks using bic. **Artificial Intelligence**, v. 260, p. 42 – 50, 2018.

CARRILLO, M. C. et al. Early risk assessment for alzheimer's disease. **Alzheimer's & Dementia**, Elsevier, v. 5, p. 182–196, 2009.

CARVALHO, V. A.; BARBOSA, M. T.; CARAMELLI, P. Brazilian version of the addenbrooke cognitive examination-revised in the diagnosis of mild alzheimer disease. **Cognitive and Behavioral Neurology**, LWW, v. 23, p. 8–13, 2010.

CHANG, T.-T.; LUNG, F.-W.; YEN, Y.-C. Depressive symptoms, cognitive impairment, and metabolic syndrome in community-dwelling elderly in southern taiwan. **Psychogeriatrics**, Wiley Online Library, v. 15, p. 109–115, 2015.

CHEN, E. Y.; DARWICHE, A.; CHOI, A. On pruning with the mdl score. **International Journal of Approximate Reasoning**, v. 92, p. 363 – 375, 2018.

CHICCO, D. Ten quick tips for machine learning in computational biology. **BioData Mining**, v. 10, n. 1, p. 35, Dec 2017.

CHICKERING, D. M. A transformational characterization of equivalent bayesian network structures. In: MORGAN KAUFMANN PUBLISHERS INC. **Proceedings of the Eleventh conference on Uncertainty in artificial intelligence**. [S.l.], 1995. p. 87–98.

_____. Optimal structure identification with greedy search. **Journal of Machine Learning Research**, JMLR, v. 3, p. 507–554, 2002.

CHICKERING, D. M. et al. **Learning Bayesian Networks is NP-hard**. [S.l.], 1994.

CHOW, C.; LIU, C. Approximating discrete probability distributions with dependence trees. **IEEE Transactions on Information Theory**, v. 14, n. 3, p. 462–467, May 1968.

CLAESKENS, G.; HJORT, N. L. et al. **Model selection and model averaging**. [S.l.]: Cambridge University Press, 2008.

COOPER, G. Causal discovery from data in the presence of selection bias. In: **Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics**. [S.l.: s.n.], 1995. p. 140–150.

COOPER, G. F. The computational complexity of probabilistic inference using bayesian belief networks. **Artificial Intelligence**, v. 42, n. 2, p. 393 – 405, 1990.

DEARBORN, J. L. et al. The metabolic syndrome and cognitive decline in the atherosclerosis risk in communities study (aric). **Dementia and geriatric cognitive disorders**, Karger Publishers, v. 38, p. 337–346, 2014.

DIK, M. G. et al. Contribution of metabolic syndrome components to cognition in older individuals. **Diabetes care**, Am Diabetes Assoc, v. 30, p. 2655–2660, 2007.

DIVERSE, P. C. G. Weight-height relationships and body mass index: some observations from the diverse populations collaboration. **American journal of physical anthropology**, v. 128, n. 1, p. 220, 2005.

EFRON, B.; TIBSHIRANI, R. J. **An introduction to the bootstrap**. [S.l.]: CRC press, 1994.

EKNOYAN, G. Adolphe quetelet (1796–1874) – the average man and indices of obesity. **Nephrology Dialysis Transplantation**, v. 23, n. 1, p. 47–51, 2008.

EXALTO, L. G. et al. The metabolic syndrome in a memory clinic population: relation with clinical profile and prognosis. **Journal of the neurological sciences**, Elsevier, v. 351, p. 18–23, 2015.

FAWCETT, T. An introduction to roc analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861–874, 2006.

FENG, L. et al. Metabolic syndrome and amnestic mild cognitive impairment: Singapore longitudinal ageing study-2 findings. **Journal of Alzheimer's Disease**, IOS Press, v. 34, p. 649–657, 2013.

FREEDMAN, D.; DIACONIS, P. On the histogram as a density estimator: L2 theory. **Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete**, Springer, v. 57, n. 4, p. 453–476, 1981.

FRIEDMAN, N.; GOLDSZMIDT, M.; WYNER, A. Data analysis with bayesian networks: A bootstrap approach. In: MORGAN KAUFMANN PUBLISHERS INC. **Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence**. [S.l.], 1999. p. 196–205.

FRIEDMAN, N. et al. Using bayesian networks to analyze expression data. **Journal of computational biology**, Mary Ann Liebert, Inc., v. 7, n. 3-4, p. 601–620, 2000.

GÁMEZ, J. A.; MATEO, J. L.; PUERTA, J. M. Learning bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood. **Data Mining and Knowledge Discovery**, v. 22, n. 1, p. 106–148, Jan 2011.

GHAHRAMANI, Z. Probabilistic machine learning and artificial intelligence. **Nature**, Nature Publishing Group, v. 521, n. 7553, p. 452, 2015.

GROSS, T. J. et al. Dependence between cognitive impairment and metabolic syndrome applied to a brazilian elderly dataset. **Artificial Intelligence in Medicine**, v. 90, p. 53 – 60, 2018.

GRUNDY, S. M. Metabolic syndrome pandemic. **Arteriosclerosis, thrombosis, and vascular biology**, Am Heart Assoc, v. 28, p. 629–636, 2008.

HAGBERG, A. A.; SCHULT, D. A.; SWART, P. J. Exploring network structure, dynamics, and function using NetworkX. In: **Proceedings of the 7th Python in Science Conference (SciPy2008)**. Pasadena, CA USA: [s.n.], 2008. p. 11–15.

HAYDUK, L. et al. Pearl's d-separation: One more step into causal thinking. **Structural Equation Modeling**, Taylor & Francis, v. 10, n. 2, p. 289–311, 2003.

HECKERMAN, D.; GEIGER, D.; CHICKERING, D. M. Learning bayesian networks: The combination of knowledge and statistical data. **Machine Learning**, v. 20, n. 3, p. 197–243, Sep 1995.

IDE, J. S.; COZMAN, F. G. Random generation of bayesian networks. In: BITTENCOURT, G.; RAMALHO, G. L. (Ed.). **Advances in Artificial Intelligence**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002. p. 366–376.

IMOTO, S. et al. Bootstrap analysis of gene networks based on bayesian networks and nonparametric regression. **Genome Informatics**, Japanese Society for Bioinformatics, v. 13, p. 369–370, 2002.

JAGER, J.; PUTNICK, D.; BORNSTEIN, M. More than just convenient: the scientific merits of homogeneous convenience samples. **Monographs of the Society for Research in Child Development**, Wiley Online Library, v. 82, n. 2, p. 13–30, 2017.

KEYS, A. et al. Indices of relative weight and obesity (*reprint*). **International Journal of Epidemiology**, v. 43, n. 3, p. 655–665, 2014.

KIM, D.; KO, S.; KANG, B. Structure learning of bayesian networks by estimation of distribution algorithms with transpose mutation. **JART**, Elsevier, v. 11, n. 4, p. 586–596, 2013.

KJAERULFF, U. B.; MADSEN, A. L. Bayesian networks and influence diagrams. **Springer Science+ Business Media**, Springer, v. 200, p. 114, 2008.

KOLLER, D.; FRIEDMAN, N. **Probabilistic graphical models: principles and techniques**. [S.l.]: MIT press, 2009.

KOMULAINEN, P. et al. Metabolic syndrome and cognitive function: a population-based follow-up study in elderly women. **Dementia and geriatric cognitive disorders**, Karger Publishers, v. 23, p. 29–34, 2007.

KREIMER, A.; HERMAN, M. A novel structure learning algorithm for optimal bayesian network: Best parents. **Procedia Computer Science**, Elsevier, v. 96, p. 43–52, 2016.

KRZYWINSKI, M.; ALTMAN, N. Points of significance: Multiple linear regression. **Nature methods**, Nature Research, v. 12, p. 1103–1104, 2015.

LAUDISIO, A. et al. Association of metabolic syndrome with cognitive function: the role of sex and age. **Clinical Nutrition**, Elsevier, v. 27, p. 747–754, 2008.

LAURITZEN, S. L.; SPIEGELHALTER, D. J. Local computations with probabilities on graphical structures and their application to expert systems. **Journal of the Royal Statistical Society. Series B (Methodological)**, [Royal Statistical Society, Wiley], v. 50, n. 2, p. 157–224, 1988.

LEE, S.-H.; YANG, K.-M.; CHO, S.-B. Integrated modular bayesian networks with selective inference for context-aware decision making. **Neurocomputing**, v. 163, p. 38 – 46, 2015.

LEVER, J.; KRZYWINSKI, M.; ALTMAN, N. Points of significance: Logistic regression. **Nature Methods**, Nature Publishing Group, v. 13, p. 541–542, 2016.

LIU, C.-L. et al. Late-life metabolic syndrome prevents cognitive decline among older men aged 75 years and over: One-year prospective cohort study. **The journal of nutrition, health & aging**, Springer, v. 17, p. 523–526, 2013.

LIU, H. et al. A new hybrid method for learning bayesian networks: Separation and reunion. **Knowledge-Based Systems**, v. 121, p. 185 – 197, 2017.

LIU, M. et al. Association between metabolic syndrome and mild cognitive impairment and its age difference in a chinese community elderly population. **Clinical endocrinology**, Wiley Online Library, v. 82, p. 844–853, 2015.

LIVINGSTON, G. et al. Dementia prevention, intervention, and care. **The Lancet**, Elsevier, v. 390, p. 2673–2734, 2017.

LUCAS, P. J.; GAAG, L. C. van der; ABU-HANNA, A. Bayesian networks in biomedicine and health-care. **Artificial Intelligence in Medicine**, v. 30, n. 3, p. 201 – 214, 2004.

MADSEN, A. L. et al. A parallel algorithm for bayesian network structure learning from large data sets. **Knowledge-Based Systems**, Elsevier, v. 117, p. 46–55, 2017.

MATTHEWS, B. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. **Biochimica et Biophysica Acta (BBA) - Protein Structure**, v. 405, n. 2, p. 442 – 451, 1975.

MIOSHI, E. et al. The addenbrooke's cognitive examination revised (ace-r): a brief cognitive test battery for dementia screening. **International journal of geriatric psychiatry**, Wiley Online Library, v. 21, p. 1078–1085, 2006.

NAGARAJAN, R.; SCUTARI, M.; LEBRE, S. **Bayesian Networks in R: With Applications in Systems Biology**. [S.l.]: Springer Verlag, 2013.

NEAPOLITAN, R. E. et al. **Learning bayesian networks**. [S.l.]: Pearson Prentice Hall Upper Saddle River, NJ, 2004. v. 38.

NEEDHAM, C. J. et al. A primer on learning in bayesian networks for computational biology. **PLoS computational biology**, Public Library of Science, v. 3, p. e129, 2007.

NG, T. P. et al. Metabolic syndrome and the risk of mild cognitive impairment and progression to dementia: follow-up of the singapore longitudinal ageing study cohort. **JAMA neurology**, American Medical Association, v. 73, p. 456–463, 2016.

OLIPHANT, T. E. Python for scientific computing. **Computing in Science and Eng.**, IEEE Computer Society, Los Alamitos, CA, USA, v. 9, p. 10–20, 2007.

_____. **Guide to NumPy**. 2. ed. [S.l.]: Trelgol Publishing USA, 2015.

PEARL, J. **Probabilistic Reasoning in Intelligent Systems**. [S.l.]: Morgan Kaufmann, 1988.

_____. **Causality**. [S.l.]: Cambridge University Press, 2009.

PEARL, J.; GLYMOUR, M.; JEWELL, N. P. **Causal inference in statistics: a primer**. [S.l.]: John Wiley & Sons, 2016.

POLIT, D. F.; BECK, C. T. **Essentials of nursing research: Appraising evidence for nursing practice**. Philadelphia, Pennsylvania: (Lippincott Williams & Wilkins, 2010.

PUGA, J. L.; KRZYWINSKI, M.; ALTMAN, N. Points of significance: Bayesian networks. **Nature methods**, Nature Research, v. 12, n. 9, p. 799–800, 2015.

REBANE, G.; PEARL, J. The recovery of causal polytrees from statistical data. In: **Proc. Workshop on Uncertainty in Artificial Intelligence**. [S.l.: s.n.], 1987. p. 222–228.

ROBINSON, R. W. Counting unlabeled acyclic digraphs. **Combinatorial mathematics V**, Springer, v. 622, n. 1977, p. 28–43, 1977.

SACHS, K. et al. Causal protein-signaling networks derived from multiparameter single-cell data. **Science**, American Association for the Advancement of Science, v. 308, n. 5721, p. 523–529, 2005.

SCHREIBER, J. M. Pomegranate: Fast and flexible probabilistic modeling in python. **Journal of Machine Learning Research**, v. 18, n. 164, p. 1–6, 2018.

SCHREIBER, J. M.; NOBLE, W. S. Finding the optimal bayesian network given a constraint graph. **PeerJ Computer Science**, PeerJ Inc., v. 3, p. e122, 2017.

SCUTARI, M.; DENIS, J.-B. **Bayesian networks: with examples in R**. [S.l.]: Chapman and Hall/CRC, 2014.

SCUTARI, M.; GRAAFLAND, C. E.; GUTIÉRREZ, J. M. Who learns better bayesian network structures: Constraint-based, score-based or hybrid algorithms? **arXiv preprint arXiv:1805.11908**, 2018.

SCUTARI, M.; NAGARAJAN, R. Identifying significant edges in graphical models of molecular networks. **Artificial Intelligence in Medicine**, Elsevier, v. 57, n. 3, p. 207–217, 2013.

SHAH, A.; WOOLF, P. Pebl: Inferring the structure of bayesian networks from knowledge and data. **JMLR**, v. 10, p. 159–162, 2009.

SIERVO, M. et al. Metabolic syndrome and longitudinal changes in cognitive function: a systematic review and meta-analysis. **Journal of Alzheimer's Disease**, IOS Press, v. 41, p. 151–161, 2014.

SILANDER, T.; KONTKANEN, P.; MYLLYMÄKI, P. On sensitivity of the map bayesian network structure to the equivalent sample size parameter. In: AUAI PRESS. **Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence**. [S.l.], 2007. p. 360–367.

STURGES, H. A. The choice of a class interval. **Journal of the American Statistical Association**, v. 21, n. 153, p. 65–66, 1926.

TOURNOY, J. et al. Association of cognitive performance with the metabolic syndrome and with glycaemia in middle-aged and older european men: the european male ageing study. **Diabetes/metabolism research and reviews**, Wiley Online Library, v. 26, p. 668–676, 2010.

TSIRLIS, K. et al. On scoring maximal ancestral graphs with the max–min hill climbing algorithm. **International Journal of Approximate Reasoning**, v. 102, p. 74 – 85, 2018.

VEMULAPALLI, V. et al. Non-obvious correlations to disease management unraveled by bayesian artificial intelligence analyses of cms data. **Artificial Intelligence in Medicine**, v. 74, p. 1 – 8, 2016.

VERMA, T.; PEARL, J. Equivalence and synthesis of causal models. In: **Proceedings of Sixth Conference on Uncertainty in Artificial Intelligence**. [S.l.: s.n.], 1991. p. 220–227.

VIEGAS, F. et al. Exploiting efficient and effective lazy semi-bayesian strategies for text classification. **Neurocomputing**, v. 307, p. 153–171, 2018.

VILLANUEVA, E.; MACIEL, C. D. Modeling associations between genetic markers using bayesian networks. **Bioinformatics**, v. 26, n. 18, p. i632–i637, 2010.

_____. Efficient methods for learning bayesian network super-structures. **Neurocomputing**, Elsevier, v. 123, p. 3–12, 2014.

WANG, J.; LIU, S. Novel binary encoding water cycle algorithm for solving bayesian network structures learning problem. **Knowledge-Based Systems**, v. 150, p. 95 – 110, 2018.

WATTS, A. S. et al. Metabolic syndrome and cognitive decline in early alzheimer's disease and healthy older adults. **Journal of Alzheimer's Disease**, IOS Press, v. 35, p. 253–265, 2013.

WHO. **World report on ageing and health**. Geneva: World Health Organization, 2015.

YANG, C. et al. Structural learning of bayesian networks by bacterial foraging optimization. **International Journal of Approximate Reasoning**, v. 69, p. 147 – 167, 2016.

YAO, T.; CHOI, A.; DARWICHE, A. Learning bayesian network parameters under equivalence constraints. **Artificial Intelligence**, v. 244, p. 239 – 257, 2017.

ZHAO, Y. et al. Learning bayesian network structures under incremental construction curricula. **Neurocomputing**, v. 258, p. 30–40, 2017.