ELIAS CÉSAR ARAÚJO DE CARVALHO

BNPA: UMA ABORDAGEM HÍBRIDA DE REDES BAYESIANAS E ANÁLISE DE TRILHAS PARA CONSTRUIR MODELOS PREDITIVOS DA ÁREA DA SAÚDE

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná, como requisito parcial obtenção do título de Doutor em Informática.

CURITIBA

2018

ELIAS CÉSAR ARAÚJO DE CARVALHO

BNPA: UMA ABORDAGEM HÍBRIDA DE REDES BAYESIANAS E ANÁLISE DE TRILHAS PARA CONSTRUIR MODELOS PREDITIVOS DA ÁREA DA SAÚDE

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná, como requisito parcial obtenção do título de Doutor em Informática.

Área de Concentração: Ciência da Computação

Orientador: Prof. Dr. Júlio César Nievola.

Coorientador: Prof. Dr. Emerson Cabrera Paraíso.

CURITIBA

2018

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central
Edilene de Oliveira dos Santos CRB 9 / 1636

C331b
2018

Carvalho, Elias César Araújo de
BNPA : uma abordagem híbrida de redes bayesianas e análise de trilhas para construir modelos preditivos da área da saúde / Elias César Araújo Carvalho ; orientador, Júilio César Nievola ; coorientador, Emerson Cabrera Paraíso. -- 2018
135 f. : il. ; 30 cm

Tese (doutorado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2018.
Bibliografia: f. 128-135

1. Informática. 2. Doença crônica. 3. Tecnologia médica. 4. Algorítmos. 5 Análise multivariada. 6. Inteligência artificial. I. Nievola, Júlio César. II. Paraíso, Ermerson Cabrera. III. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática. IV. Título

CDD 20. ed. – 004

Pontifícia Universidade Católica do Paraná
Escola Politécnica
Programa de Pós-Graduação em Informática

PUCPR
GRUPO MARISTA

ATA DE SESSÃO PÚBLICA

DEFESA DE TESE DE DOUTORADO Nº 57/2018

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA – PPGIa
PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ - PUCPR

Em sessão pública realizada às 08h30 de 29 de Maio de 2018, no Auditório Guglielmo Marconi – Bloco 8 – Térreo, ocorreu a defesa da tese de doutorado intitulada **"BNPA: Uma Abordagem Híbrida de Redes Bayesianas e Análise de Trilhas para Construir Modelos Preditivos da área da Saúde"** elaborada pelo aluno **Elias Cesar A. Carvalho**, como requisito parcial para a obtenção do título de **Doutor em Informática**, na área de concentração **Ciência da Computação,** perante a banca examinadora composta pelos seguintes membros:

Prof. Dr. Julio Cesar Nievola (orientador) - PUCPR

Prof. Dr. Emerson Cabrera Paraiso – PUCPR

Prof. Dr. Alceu de Souza Britto Junior – PUCPR

Prof. Dr. Roberto Tadeu Raittz - UFPR

Prof.ª Dr.ª Deborah Ribeiro Carvalho – PUCPR/PPGTS

Após a apresentação da tese pelo aluno e correspondente arguição, a banca examinadora emitiu o seguinte parecer sobre a tese:

Membro	Parecer	
Prof. Dr. Julio Cesar Nievola	(X) Aprovada	( ) Reprovada
Prof. Dr. Emerson Cabrera Paraiso	(X) Aprovada	( ) Reprovada
Prof. Dr. Alceu de Souza Britto Junior	(X) Aprovada	( ) Reprovada
Prof. Dr. Roberto Tadeu Raittz	(X) Aprovada	( ) Reprovada
Prof.ª Dr.ª Deborah Ribeiro Carvalho	(X) Aprovada	( ) Reprovada

Portanto, conforme as normas regimentais do PPGIa e da PUCPR, a tese foi considerada:

( ) **APROVADO**
(aprovação condicionada ao atendimento integral das correções e melhorias recomendadas pela banca examinadora, conforme anexo, dentro do prazo regimental)

( ) **REPROVADO**

E, para constar, lavrou-se a presente ata que vai assinada por todos os membros da banca examinadora.
Curitiba, 29 de Maio de 2018.

Prof. Dr. Julio Cesar Nievola

Prof. Dr. Emerson Cabrera Paraiso

Prof. Dr. Alceu de Souza Britto Junior

Prof. Dr. Roberto Tadeu Raittz

Prof.ª Dr.ª Deborah Ribeiro Carvalho

Programa de Pós-Graduação em Informática
PPGIa
PUCPR

Rua Imaculada Conceição, 1155 Prado Velho CEP 80215 901 Curitiba Paraná Brasil
Tel.: (41) 3271 1669 Fax.: (41) 3271 2121 www.ppgia.pucpr.br

Dedico este trabalho a minha família e a DEUS,
sem os quais eu não teria nem força nem motivação para continuar.

Felix, qui potuit rerum cognoscere causas" (Virgil 29 BC)

"Abençoado aquele que conseguiu entender a causa das coisas" (Virgil 29 BC)

AGRADECIMENTOS

Primeiramente a Deus, porque sem ele nada seria possível.

Ao meu orientador Prof. Dr. Júlio César Nievola, por todos os ensinamentos e pela orientação segura.

Ao Prof. Dr. Emerson Cabrera Paraiso pelas orientações e importantes contribuiçõesdadas à realização deste trabalho.

Aos meus pais Raimundo e Maria José.

À minha querida esposa Adélia.

A meus filhos Gustavo e Natália.

À UEM pelo apoio financeiro recebido durante a realização deste trabalho.

Ao PPGIa pela oportunidade e suporte oferecidos ao desenvolvimento destetrabalho.

A CAPES pelo apoio financeiro.

A todos que direta ou indiretamente colaboraram na execução deste trabalho.

RESUMO

Compreender como as doenças crônicas não transmissíveis (DCNT), como por exemplo as doenças que afetam o coração, se desenvolvem é um dos desafios atuais da epidemiologia. O processo para alcançar essa compreensão é complexo pois requer a análise de todas as variáveis do processo e suas possíveis interações de forma simultânea. Neste sentido, se acredita que o método de análise de trilhas (AT), que permite estimar relações multivariadas com a perspectiva de trajetórias representando o modelo por meio de grafos acíclicos dirigidos (GAD), seja adequado para este tipo de análise. Além de permitir análises mais complexas e modelar situações mais próximas da realidade, AT também possui métodos de estimação já consolidados para cada tipo de variável e permite estimar efeitos diretos e indiretos. No entanto, o método de AT não tem capacidade de inferir causalidades a partir de dados, a construção de modelos neste caso é dependente de especialistas. Essa tarefa pode se tornar penosa frente a uma grande quantidade de variáveis e maior complexidade do sistema. Em busca por soluções que preencham essa lacuna, a literatura mostrou que a metodologia de aprendizagem de estrutura de redes Bayesianas (RB) apresenta diversos algoritmos, já bem consolidados, o quais permitem que se aprenda a estrutura de uma RB a partir de um conjunto de dados. Neste contexto, esta tese tem a proposta de propor e avaliar um método computacional que habilite a metodologia de AT a inferir causalidades a partir de um conjunto de dados e gerar o modelo de AT. Este método une a capacidade da técnica de RB em inferir causalidades a partir de dados à robustez estatística da técnica de AT dando origem ao nome *bnpa*, do inglês, *Bayesian Networks & Path Analysis*. Foram realizados experimentos com o conjunto de dados do *Canadian Community Health Survey* (CCHS) que é composto por 1381 variáveis e 124.929 registros e contém informações sobre pacientes com/sem doença cardiovascular (DCV). O conjunto foi pré-processado, restando 63.884 registros. Dois especialistas doutores com experiência em pesquisa clínica e cardiologia selecionaram 14 variáveis para serem incluídas no estudo. O método *bnpa* utilizou 4 algoritmos de de aprendizagem de estrutura de RBs baseados em restrição e mais 2 algoritmos baseados em pontuação para a aprendizagem da estrutura das RBs a partir do conjunto de dados com seus respectivos testes, gerando 10 RBs. O mesmos especialistas que selecionaram as variáveis, avaliaram as RBs aprendidas com base em critérios específicos e selecionaram a RB que melhor representa a causalidade do estudo. O método *bnpa* gerou a partir dessa estrutura o grafo de AT, os índices de avaliação de qualidade do ajuste e a matriz de correlação residual. O grafo de AT foi avaliado e os índices de qualidade do ajuste ficaram dentro do valor de corte estabelecido pela literatura sinalizando um bom ajuste do modelo aos dados. A matriz de correlação residual apresentou valores abaixo de 0.10 confirmando um bom desempenho do modelo. Esses índices foram comparados a índices de estudos similares indicando desempenho similar e melhor em alguns casos. Concluiu-se que a combinação BN x AT pode ser um avanço no sentido de capacitar a técnica de AT com habilidades de inferir causalidades a partir de dados para então criar o modelo de AT e gerar inferências, confirmando a hipótese desta tese. Todas as funcionalidades do método proposto foram implementadas no pacote R *bnpa* (*Bayesian Networks Path Analysis*), o qual está disponível para *download*, publicado no site do projeto R em: https://cran.r-project.org/web/packages.

Palavras-chave: Modelos Causais. Redes Bayesianas. Análise de Trilhas. Aprendizagem de Estrutura. Mediação de Modelos. Índices de Ajuste. Matriz de Correlação Residual.

ABSTRACT

Understanding how chronic noncommunicable diseases (CNCDs), such as diseases affecting the heart, are growing is one of the current challenges of epidemiology. The process for achieving this understanding is complex because it requires the analysis of all process variables and their possible interactions simultaneously. In this sense, it is believed that the method of path analysis (PA), which allows the estimation of multivariate relationships with the perspective of trajectories representing the model using directed acyclic graphs (DAGs), is adequate for this type of analysis. In addition to allowing more complex analysis and modeling situations closer to reality, PA has already consolidated estimation methods for each type of variable and allows estimation of direct and indirect effects. However, the PA method has no ability to infer causalities from data; the construction of models in this case is dependent on experts. This task can become difficult in the face of a large number of variables and greater system complexity. In the search for solutions that fill this gap, the literature has shown that Bayesian networks (BN) structure learning methodology presents several well-consolidated algorithms, which allow one to learn the structure of an RB from a set of data. In this context, this thesis proposes and evaluates a computational method that enables the PA methodology to infer causality from a dataset and generate the PA model. This method unites the ability of the BN technique to infer causalities from data to the statistical robustness of the PA technique giving rise to the name *bnpa* from the english Bayesian Networks & Path Analysis. Experiments were conducted with the Canadian Community Health Survey (CCHS) dataset, which consists of 1381 variables and 124,929 records and contains information on patients with / without cardiovascular disease (CVD). The dataset was preprocessed, leaving 63,884 records remaining. Two specialist doctors with experience in clinical research and cardiology selected 14 variables to be included in the study. The *bnpa* method used 4 constraint-based algorithms and 2 score-based for learning the structure of RBs from the dataset with their respective tests, generating 10 RBs. The same experts who selected the variables, evaluated the RBs learned based on specific criteria and selected the RB that best represents the causality of the study. The *bnpa* method generated from this structure the PA graph, the indices of quality for evaluation of the fit and the residual correlation matrix. The PA graph was evaluated and the adjustment quality indexes were within the cut-off value established by the literature signaling a good fit of the model to the data. The residual correlation matrix presented values below 0.10 confirming a good performance of the model. These indices were compared to similar study indices indicating similar and better performance in some cases. It was concluded that the combination BN x PA can be an advance in order to enable the PA technique with the ability to infer causalities from data to create the PA model and generate inferences, confirming the hypothesis of this thesis. All the functionalities of the proposed method were implemented in the R *bnpa* (Bayesian Networks & Path Analysis) package, which is available for download, published on project site R at: https://cran.r-project.org/web/packages.

Keywords: Causal Models. Bayesian Networks. Path Analysis. Structure Learning. Mediation of Models. Adjustment Indices. Matrix of Residual Correlation.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AAERB	-Algoritmos de Aprendizagem de Estrutura de Redes Bayesianas
ADF	-*Asymptotically Distribution-Free*
AFC	-Análise Fatorial Combinatória
AGFI	-*Adjusted Goodness-of-fit*
AIC	-*Akaike Information Criteria*
AINS-BN	-*Adaptive Importance Sampling Bayesian Networks*
AT	-Análise de Trajetórias
BCC	-*Browne-Cudeck Criterion*
BENEDICT	-*BElief NEtworks DIscovery using Cut-set Techniques*
BDE	-*Bayesian Dirichlet equivalent score*
BIC	-*Bayes Information Criterion*
BNPA	-*Bayesian Networks and Path Analysis*
CAM	-Cobertura Aproximada de Markov
CCHS	-*Canadian Community Health Survey*
CFI	-*Comparative Fit Index*
CM	-Cobertura de Markov
CN	-*Critical N*
DCC	- Doença Cardíaca Coronariana
DCNT	- Doenças Crônicas Não Transmissíveis
DCV	- Doença Cárdio Vascular
DP	-Desvio Padrão
DPC	-Distribuição de Probabilidade Conjunta
ECVI	-*Expected Cross-Validation Index*
EM	-Estimation-Maximization
FAST-IAMB	-*Fast Incremental Association*
FCS	-*Fully Conditional Specification*
GAD	-Gráficos Acíclicos Dirigidos
GFI	-*Goodness-of- Fit Index*
GL	-Graus de Liberdade
GLS	-*Generalized Least Squares*
GS	-*Grow-Shrink*
HC	-*Hill Climbing*

IA	-Inteligência Artificial
IAMB	-*Incremental Association Markov Blanket*
IC	-Independência Condicional
INTER-IAMB	-*Interleaved Incremental Association*
IR	-Imposto de Renda
MD	-Mineração de Dados
MDL	-Minimal Description Length
MEE	-Modelagem de Equações Estruturais
MLE	-*Maximum Likelihood Estimation*
MMHC	-*Max-Min Hill Climbing*
MMPC	-*Max-Min Parent Children*
NCP	-*Noncentrality Parameter*
NFI	-*Normed Fit Index*
NNFI	-*Nor-Normed Fit Index*
PA	-*Path Analysis*
PAS	- Pressão Arterial Sistólica
PC	- Peter and Clarkl
PCFI	-*Pasimony CFI*
PGFI	-*Parsimony Goodness-of-fit Index*
PNFI	-*Parsimony NFI*
PRATIO	-*Parsimony Ratio*
RBs	-Redes Bayesianas
RBGs	-Redes Bayesianas Gaussianas
RFI	-*Relative Fit Index*
RMR	-*Root Mean Square Residual*
RMSEA	-*Root Mean Square Error of Approximation*
RSMAX2	-*General 2-Phase Restricted Maximization*
SGS	- Spirtes, Glymour, and Scheines
SI-HITON-PC	-*Semi-Interleaved Hiton-PC*
SLA	-*Simple Learning Algorithm*
SMILE	-*Structural Modelling, Inference, and Learning Engine*
SRA	- *Search-space Reduction Algorithm*
SRMR	-*Standardized Root Mean Square Residual*
TABU	-*Tabu Search*

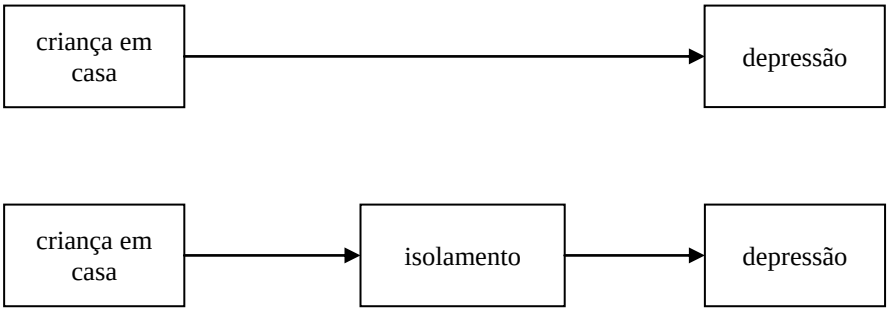
TPC	-Tabela de Probabilidades Condicionais
TPDA	-*Three-Phase Dependency Analysis Algorithm*
TLI	-*Tucker-Lewis Index*
VD	-Variável Dependente
VI	-Variável Independente
VEX	-Variáveis Exógenas
VEN	-Variáveis Endógenas
VI	-Variável Independente
WLS	-*Weighted Least Squares*

SUMÁRIO

1 INTRODUÇÃO

Epidemiologia de curso de vida (ECV) é uma área de pesquisa que busca a compreensão de como as doenças crônicas não transmissíveis (DCNT) se desenvolvem (KUH e SHLOMO, 2004). O processo para chegar a essa compreensão pode se tornar complexo, e neste sentido, os métodos estatísticos mais comumente utilizados na epidemiologia podem não ser suficientes para realizar determinada análise de dados (GAMBORG et al., 2011). Um primeiro exemplo dessa complexidade visa um estudo onde se deseja avaliar o processo que leva o tamanho do corpo à incidência de doença cardíaca coronariana (DCC). Esse estudo requer um mecanismo de investigação mais profundo pelo fato de que o tamanho do corpo geralmente também está associado a diversos outros fatores de risco para DCC, como por exemplo, pressão arterial sistólica (PAS) (GAMBORG et al., 2011). Um segundo exemplo, objetiva comparar os resultados de uma medicação com os resultados de uma terapia cognitiva-comportamental e verificar qual das duas alternativas apresenta melhor resultado em reduzir os sintomas de depressão (STREINER, 2005). A primeira hipótese, baseada na literatura, dita que mulheres casadas com crianças jovens em casa têm mais depressão do que mulheres de mesma idade e estado civil, que não têm filhos jovens em casa, ou seja, crianças provocam a depressão dessas mulheres. No entanto, o mundo real onde se vive cria situações muito mais complicadas do que essa. Neste contexto, uma segunda hipótese dita que ficar em casa leva ao isolamento e este leva disforia, um mal-estar psíquico acompanhado por depressão, ansiedade, tristeza, melancolia e pessimismo. A Figura 1 apresenta dois modelos causais representando essas situações.

Figura 1 - Dois possíveis modelos de efeitos de crianças em casa sobre depressão



Fonte: STREINER, 2005, traduzida.

Nos dois modelos fica óbvio que a variável “criança em casa” contribui como variável independente (VI) para a variável dependente (VD) “depressão”. No segundo modelo surge uma dúvida sobre a variável “isolamento”. É neste ponto que o modelo começa a se complicar, porque esta variável é considerada VI com relação a variável “depressão” e VD

com relação a "criança em casa". Então surge a dúvida: Como se deve analisar este modelo? O problema citado se torna ainda mais complexo conforme aumenta a quantidade de variáveis, pois outros fatores podem influenciar a depressão, como por exemplo: mudanças hormonais, depressões anteriores, história de depressão na família, *stress*, causado pelas crianças em casa, que por sua vez gera mal humor e outros possíveis resultados além da disforia podem surgir.

Os exemplos citados e os questionamentos apresentados mostram que considerando a epidemiologia, existe a necessidade promover métodos estatísticos para expandir a competência explicativa dos pesquisadores e a eficiência dos resultados gerados. Esse é um dos principais objetivos das técnicas multivariadas estatísticas, porém entre essas técnicas, como a regressão múltipla, análise fatorial, análise multivariada de variância, análise discriminante, análise canônica e outras, existe a limitação de examinar somente uma relação por vez, ou seja, mesmo aquelas que permitem múltiplas VDs, só permitem analisar o relacionamento entre várias VIs e apenas uma VD (HAIR, 2005). Dentro deste contexto, quando o objetivo for observar e analisar todas essas variáveis e suas possíveis/diversas interações simultaneamente, estratégias que permitam analisar modelos mais complexos e realistas do que outros métodos estatísticos são requeridos. Neste caso, uso da análise de trajetórias (AT) ou *path analysis* (PA) torna-se adequado. Entenda-se por "adequado...", um método capaz de permitir que o efeito de um fator de risco para uma doença seja decomposto em efeitos indiretos mediados por outras variáveis do modelo e em efeitos diretos não mediados. Esta decomposição tem o potencial de aprimorar a compreensão dos mecanismos por trás dos fatores de riscos que contribuem ou não para o desenvolvimento de doenças (GAMBORG et al., 2011). Neste sentido, entende-se que o método de AT tem a capacidade para preencher esta lacuna, pois permite a estimativa de relações multivariadas com a perspectiva de trajetórias por meio de grafos acíclicos dirigidos (GADs) (HAIR, 2005; MAROCO, 2010, BEAUJEAN, 2014; KLINE, 2015).

Durante a revisão da literatura para a elaboração desta tese, se constatou que o método de AT apesar de apresentar métodos estatísticos robustos e possuir características que o capacitem a analisar modelos complexos, este não tem a capacidade de provar causalidade (WRIGHT, 1934; STREINER, 2005; BEAUJEAN, 2014; KLINE, 2015). Não menos importante, uma recente publicação (ZHANG, 2017) destacou que o uso AT é limitado no contexto da pesquisa clínica. Essa constatação foi justificada pela provável dificuldade técnica em usar o método e na complexidade de identificar um modelo coerente estatisticamente e

teoricamente, um problema que se torna exponencial concomitante com o aumento das variáveis do modelo.

1.1 OBJETIVOS

Esta tese tem como objetivo geral propor e avaliar um método computacional que forneça suporte para criação de modelos causais utilizando a combinação de técnicas de modelagem de redes Bayesianas (RBs) e técnicas de modelagem de análise de trilhas (ATs).

Como objetivos específicos se destacam:

1. criar um método computacional que permita identificar os tipos de variáveis do conjunto de dados a ser tratado;
2. utilizar métodos adicionais para auxiliar o pré-processamento de dados como identificação e tratamento de dados faltantes, valores anormais (*outliers*), e multicolinearidade;
3. desenvolver um método computacional que supere a falha dos métodos de criação de RBs em não identificar variáveis tipicamente preditoras e variáveis tipicamente de desfecho;
4. propor um método computacional híbrido que utilize as técnicas de RBs e ATs em conjunto para a criação de modelos causais;
5. avaliar e validar o método proposto, analisando os resultados obtidos;
6. por fim, que este método possua meios que facilitem o uso de AT por pesquisadores clínicos.

1.2 JUSTIFICATIVA

O método de AT possibilita analisar todas as variáveis e suas possíveis interações simultaneamente, além disso, se comparados a modelos tradicionais, com exceção de modelos de séries temporais, os modelos de AT também apresentam as seguintes características (MAROCO, 2010; BEAUJEAN, 2014; KLINE, 2015): a) possui métodos de estimação já consolidados para cada tipo de variável (dicotômica, ordinal, contagem e contínua); 2) permite modelar e estimar os efeitos diretos e indiretos e; 3) possibilita o uso de variáveis instrumentais e a modelagem dos erros, caso seja o desejo do pesquisador. Como vantagem

final descata-se a capacidade de AT em representar o modelo por meio de um GAD e o fato de este ser um método adotado cada vez mais, embora ainda de forma limitada, na área de epidemiologia como ferramenta auxiliar na inferência causal (GREENLAND, PEARL e ROBINS, 1999; ROBINS, HERNAN e BRUMBACK, 2000; FUCHS, 2006; VANDERWEELE, VANSTEELANDT e ROBINS, 2010; RICHMOND et al., 2014; BROADBENT, 2015; VANDERBROUCKE, BROADBENT e PEARCE, 2016). Neste caso, o GAD atua como uma ferramenta conveniente para auxiliar os pesquisadores a comunicar suas suposições, defender hipóteses e direcionar análises. O método de AT ainda tem a proposta de fornecer estimativas de efeitos causais sob suposições mais coerentes do que os projetos convencionais de pesquisas epidemiológicas (GLYMOUR; KUBZANSKY, 2017).

Apesar de todas as vantagens observadas, como destacado na introdução desta tese, identificou-se duas lacunas que o método AT não preenche: a) AT não tem capacidade de identificar causalidades (WRIGHT, 1934; STREINER, 2005; BEAUJEAN, 2014; KLINE, 2015) e b) AT ainda é pouco usado no contexto da pesquisa clínica (ZHANG, 2017), considerando o montante de projetos de pesquisa e publicações da área.

Em busca por soluções que possam auxiliar no preenchimento da primeira lacuna a revisão da literatura mostrou que métodos de aprendizagem de estruturas de redes Bayesianas possuem diversos algoritmos bem consolidados com capacidade de identificar relacionamentos entre variáveis e aprender GADs, (Chow e Liu, 1968; Srinivas, 1990; Spyrtes e Glymour, 1993; Cheng, Bell e Liu 1997a; Cheng, Bell e Liu 1997b; Spyrtes, Glymour e Scheines, 2000; Margaritis, 2003; Tsamardinos, 2003; Yaramakala e Margaritis, 2003; Tsamardinos, 2006). Além disso, métodos de RBs possuem a falha de não mensurar efeitos indiretos, neste caso o método computacional proposto por esta tese também irá preencher, embora de forma indireta, esta lacuna. A segunda lacuna será preenchida com funcionalidades implementadas no método *bnpa* as quais visam facilitar o uso do método de AT por parte de pesquisadores clínicos.

1.3 INEDITISMO DO TRABALHO

Na revisão da literatura, apresentada no capítulo 3, se encontrou diversas abordagens usando técnicas de modelagem que implementam modelagem de equações estruturais (MEE) (AT é uma subárea de MEE), ATs e RBs. No caso de artigos que usam MEE e AT fica claro que esses métodos não identificaram causalidades, pois estes foram utilizados para identificar ou não um bom ajuste dos dados ao modelo. Na maioria dos artigos, esses métodos foram

usados de forma independente para construir modelos preditivos e ajudar na prevenção de doenças. Além disso, os modelos foram construídos com a ajuda de especialistas, provavelmente estatísticos e especialistas em RBs, ATs e/ou MEE, os quais podem ser caros e demorados. Destes, nenhum estudo utilizou algoritmos para aprender modelos preditivos a partir de dados conforme proposto nesta tese.

De acordo com o nosso conhecimento, este estudo é o primero a propor a construção de modelos de AT a partir de um conjunto de dados. Também destacamos que é o único estudo que utiliza mais de um algoritmo de aprendizagem de estruturade RBs (AAERBs) para construção de um modelo de AT. Por fim, este estudo é ímpar no sentido de fornecer um algoritmo que prepara variáveis para atuar como tipicamente preditoras ou tipicamente de desfecho durante a aprendizagem da estrutura da RB e outro algoritmo para criar a partir da estrutura da RB um modelo de entrada para criação do modelo de AT.

1.4 MOTIVAÇÃO

A compreensão dos mecanismos por trás dos fatores de riscos que contribuem ou não para o desenvolvimento de doenças ainda é um problema desafiador na área da epidemiologia. Publicações com análises estatísticas duvidosas motivam a busca por métodos estatísticos mais robustos. É comum encontrar em artigos científicos frases sobre melhorias significativas como: “houve uma melhoria significativa no funcionamento cognitivo”. Neste caso, o que o quer dizer a palavra “significativa”? Quer dizer que a melhoria foi grande, importante ou que o 'valor-p' foi inferior a 0,05? Um 'valor-p' pequeno não garante que a melhoria seja grande ou importante (ESPÍRITO SANTO, 2017).

Portanto, desenvolver métodos computacionais que contribuam com expansão da capacidade dos pesquisadores em explicar esses mecanismos e melhorar a eficiência dos resultados gerados ainda é necessário. Neste contexto, o método de AT fornece uma robusta base estatística com métodos de estimação já consolidados para cada tipo de variável e permite modelar e estimar efeitos indiretos, parece ser um método promissor. AT é particularmente útil quando o pesquisador deseja analisar situações em que ocorrem relações simultâneas, ou seja, uma VD se torna VI em relações posteriores de dependência e vice-versa. Neste caso, modelagem por AT possibilita examinar uma série de relações de dependência simultaneamente (HAIR et al., 1998).

Portanto, motiva o desenvolvimento deste estudo a idéia de que capacitar o método de AT a inferir causalidades a partir de dados contribuirá com o trabalho dos pesquisadores, pois

o processo para criar um modelo de AT estatisticamente e teoricamente coerente é complexo. Essa complexidade vai aumentando de forma exponencial conforme aumenta o número de variáveis tornando a tarefa do pesquisador ainda mais difícil. Dessa forma, a criação do método proposto por esta tese visa superar essa complexidade aprendendo a estrutura do modelo de entrada para criar o modelo de AT a partir dos dados.

O pouco uso do método de AT no contexto da pesquisa clínica se comparado aos tradicionais métodos estatísticos também motiva a implementação do método *bnpa*, pois neste foram implementados métodos que facilitam o seu uso por pesquisadores clínicos.

1.5 CONTRIBUIÇÕES

Pode-se destacar como contribuição desta tese:

1. o fornecimento de dois algoritmos: o primeiro ajudará os algoritmos AAERBs a identificar e tratar as variáveis tipicamente preditoras (VIs) e tipicamente de desfecho (VDs), um procedimento essencial em pesquisa clínica, e o segundo algoritmo auxiliará na automatização da criação do modelo de entrada de AT se baseando na estrutura de RB aprendida;
2. a geração de informações (tabelas, gráficos e índices) para ajudar os pesquisadores a avaliar a qualidade do(s) modelo(s) de AT criado(s);
3. funções adicionais, como identificação e tratamento de *outliers*, verificação automática de tipo de variáveis, entre outras foram criadas para o método de modo a facilitar a vida dos pesquisadores e poderão ser utilizadas a parte se necessário;
4. avaliação de forma semi-automática da melhor estrutura de RB aprendida usando validação cruzada;
5. capacitação do método de AT em identificar causalidades por meio de algoritmos de AAERBs para geração automática do modelo de entrada de AT;
6. disponibilização do método proposto, no *CRAN-R*, uma rede de servidores *ftp* e *web* em todo o mundo que armazena versões de códigos (chamados de pacotes) e documentação atualizadas para o ambiente R.

1.6 HIPÓTESE DE PESQUISA

A hipótese básica desta tese, comprovada por meio dos resultados apresentados em experimentos realizados pelo método proposto, é de que é possível capacitar o método de AT a inferir causalidades a partir de dados utilizando AAERBs e gerar um modelo de entrada para criar modelos de ATs com boa qualidade de ajuste aos dados.

1.7 ESTRUTURA DA TESE

Esta tese está organizada em 06 (seis) capítulos. O Capítulo 2 apresenta pressupostos teóricos sobre inferência causal estatística, modelos gráficos, redes Bayesianas, análise de trajetórias, neste capítulo se apresenta conceitos importantes para inferência causal, destacam-se também os algoritmos de aprendizagem de RBs, descreve-se como são criados modelos de AT ressaltando as formas de avaliar a qualidade de ajuste do modelo. No Capítulo 3 se apresentam trabalhos correlatos à área de estudo, destacando-se como foram criados os modelos, os métodos utilizados e como foram avaliados. No capítulo 4 se detalha o método proposto para esta tese, que inclui informações sobre métodos de pré-processamento de dados, construção da RB, criação e validação do modelo de AT. O Capítulo 5 aborda a análise e discussão dos resultados. E finalmente, no Capítulo 6 é apresentado a discussão dos resultados, as conclusões e trabalhos futuros.

2 PRESSUPOSTOS TEÓRICOS

2.1 CONSIDERAÇÕES INICIAIS

O objetivo deste capítulo é apresentar os principais conceitos que serão utilizados por esta tese. Na seção 2.2 serão abordados os conceitos sobre inferência causal estatística, na seção 2.3 serão apresentados modelos gráficos e suas aplicações, na seção 2.4 conceitua-se as redes Bayesianas, na seção 2.5 A análise de trilhas é explorada.

2.2 INFERÊNCIA CAUSAL ESTATÍSTICA

É fato que nos dias atuais armazena-se cada vez mais dados, provenientes das mais diversas fontes. Neste contexto, gerou-se a necessidade de dar sentido a esses dados de forma que eles possam guiar a tomada de decisões e criação de políticas por partes dos gestores de diversas áreas. A partir deste fato surgiu a necessidade de estudar a causalidade, um tópico separado da estatística, que ao ser abordado com rigor só vem a enriquecer os métodos estatísticos. Portanto, a causalidade não deve ser encarada como um concorrente da estatística, mas sim um método capaz de descobrir fatos relacionados ao funcionamento do mundo real, os quais não poderiam ser descobertos por métodos tradicionais (PEARL; GLYMOUR; JEWELL, 2016).

2.2.1 O Paradoxo de Simpson

Esta seção destina-se a esclarecer que ao se executar uma análise estatística não basta apenas aplicar os métodos existentes, deve-se também conhecer a história por trás dos dados.

Um dos enigmas mais intrigantes da literatura estatística é o Paradoxo de Simpson Simpson (1951). Este paradoxo refere-se a inversão da associação estatística de uma determinada população quando os dados são analisados para uma subpopulação. Em 1951, Simpson avaliou pacientes portadores de uma doença que tinham a opção de experimentar um novo medicamento. O resultado do experimento, apresentados no Quadro1, demonstrou que considerando todos os pacientes, uma quantidade menor se recuperou se comparados aos que não tomaram o medicamento. Neste mesmo experimento, ao considerar o sexo, o estudo identificou a recuperação de mais pacientes que tomaram o medicamento quando comparados àqueles que não o tomaram. Diante do exposto, gerou-se a impressão de que o medicamento

tem maior efeito curativo sobre homens e mulheres, e ao mesmo tempo, este piora a situação da população. Por esse motivo considera-se essa situação como um paradoxo. Com este resultado, como deve ser prescrito este medicamento por um médico? No caso de um gestor da saúde pública, como este deve avaliar a eficácia do medicamento?

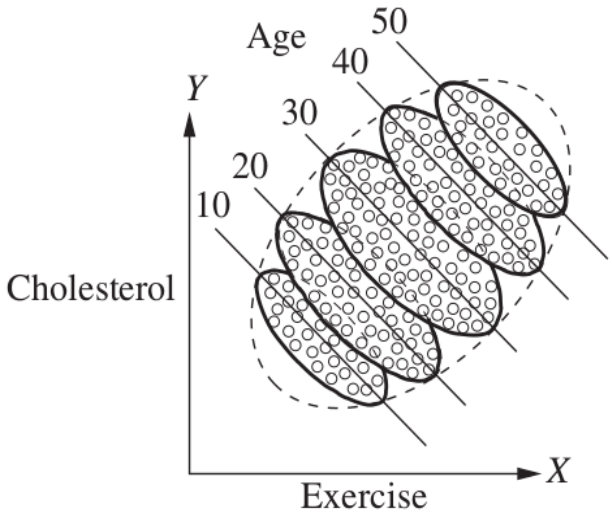
Quadro 1 - Resultados de um estudo sobre um novo medicamento, considerando o sexo

Pacientes	**Com medicamento**	**Sem medicamento**
Homens	81 de 87 recuperados (93%)	234 de 270 recuperados (87%)
Mulheres	192 de 263 recuperados (73%)	55 de 80 recuperados (69%)
Todos	273 de 350 recuperados (78%)	289 de 350 recuperados (83%)

Fonte: PEARL; GLYMOUR; JEWELL 2016.

Essas duas perguntas são difíceis de responder por meio de estatísticas simples. É necessário conhecer a história por trás dos dados, ou seja, deve-se obter informações adicionais para identificar o mecanismo causal que provocou o resultado que foi observado (PEARL; GLYMOUR; JEWELL, 2016). Por exemplo, dois novos fatores podem ser adicionados ao estudo: 1) A variável "estrogênio", um hormônio feminino que gera um efeito negativo na recuperação de mulheres, tomando ou não o novo medicamento e 2) Neste estudo as mulheres são mais propensas a tomar o medicamento se comparadas aos homens. Como consequência disso, ao selecionar um paciente que fez uso do medicamento, há maiores chances deste ser uma mulher e consequentemente ter menos chance de ter se recuperado com sucesso. Isso leva a conclusão de que para avaliar o medicamento de forma mais eficiente, é necessário comparar indivíduos do mesmo sexo. Este conceito é conhecido como "dados segregados" e, neste caso, tem o objetivo de evitar a interferência de qualquer fator causal não pertinente a determinado sexo. Em um outro exemplo, considerando-se um estudo que mede o exercício semanal e o nível de colesterol em grupos de cinco faixas etárias (FIGURA 2). Plotando-se a prática de exercícios no eixo *X* e o nível de colesterol no eixo *Y* e segregando por idade, observa-se que todos os grupos etários tendem para baixo.

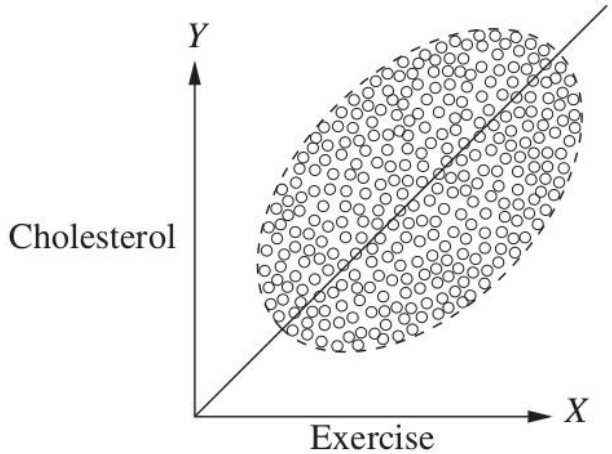
Figura 2 – Estudo sobre exercícios x níveis de colesterol segregado por idade



Fonte: PEARL; GLYMOUR; JEWELL, 2016.

Se, no entanto, for usado o mesmo gráfico de dispersão, mas sem segregar por gênero (FIGURA 2), observa-se uma tendência geral para cima, ou seja, quanto mais exercícios uma pessoa pratica, maior é o seu nível de colesterol. Novamente, para resolver este problema, deve-se conhecer a história por trás dos dados. Se sabemos que as pessoas mais velhas são mais propensas a se exercitar (FIGURA 3) e também que são mais propensas a ter colesterol alto, independentemente do exercício, então a reversão é facilmente explicada. Neste caso a análise da causalidade deve ser feita de modo que se compare pessoas de mesma idade apenas.

Figura 3 - Resultados do estudo sobre exercícios x níveis de colesterol, não-agregados



Fonte: PEARL; GLYMOUR; JEWELL, 2016.

Como a estatística tradicionalmente adverte: “correlação não é causalidade”, portanto ao se avaliar a causalidade não se deve usar apenas a correlação, outros métodos são requeridos. A próxima seção detalha os conceitos e formulações sobre correlação, variância e covariância.

2.2.2 Análise de correlação

Análise de correlação consiste em um método estatístico que fornecerá um valor, conhecido como coeficiente de correlação. Este coeficiente refere a relação entre duas variáveis ou a ausência dela e indica aintensidade da variação conjunta entre duas variáveis. Essa variação pode ser linear, ou seja, a mudança de uma variável gera uma mudança constante no valor de outra variável ou não. Esta variação ainda pode ser positiva quando mudança é similar entre as duas variáveis (aumentam ou diminuem juntas) ou negativa (quando uma aumenta a outra diminui). Por fim, o coeficiente de correlação pode variar entre os valores absolutos 0 e 1, quanto mais próximo de 1, mais forte é a correlação entre as variáveis. Este coeficiente ainda pode ser zero indicando falta de relação linear (BUSSAB; MORETTIN, 2017).

A fim de evitar a ocorrência de relações espúrias, antes de executar uma análise de correlação, o pesquisador deve analisar o conjunto de dados em busca de valores anormais, tambéo conhecido como *outliers*e eliminá-los. Esses valores são representados por um valor atípico representando uma grande diferença dos demais valores da série). Dados com valores anormais podem comprometer fortemente os valores do coeficiente de correlação, induzindo o pesquisador a cometer erros do tipo I (rejeitar uma hipótese quando esta é verdadeira) e erros do tipo II (aceitar uma hipótese falsa) (OSBORNE e WATERS, 2002). Além disso, é necessário verificar a independência das observações, ou seja, a influência causal de X_1 para Y_1 não deve alterar a influência causal de X_2 para Y_2. A correlação é somente uma medida de associação e portanto não permite qualquer conclusão sobre causa e efeito. Não é possível fazer inferências sobre causalidades com base na correlação.

Correlação não implica necessariamente causalidade

"*Correlação não implica necessariamente em causalidade*", esta frase comumente utilizada em livros e cursos de estatística, enfatiza que, a existência de uma correlação entre duas variáveis não implica necessariamente que uma causa a outra (TSHILIDZI, 2015). O

correto entendimento da causalidade das coisas é uma preocupação da humanidade desde os tempos de Aristóteles. Na área médica, foco desta tese, a premissa básica é a de que alguns medicamentos curam certas doenças. No entanto, a recomendação de uma medicação para a cura de uma doença requer um bom entendimento da causalidade, mesmo que esta ofereça uma solução sub-ótima. E para entender a causalidade é necessário entender os princípios da correlação (TSHILIDZI, 2015). Alguns exemplos que comprovam que correlação não implica necessariamente em causalidade são descritos a seguir.

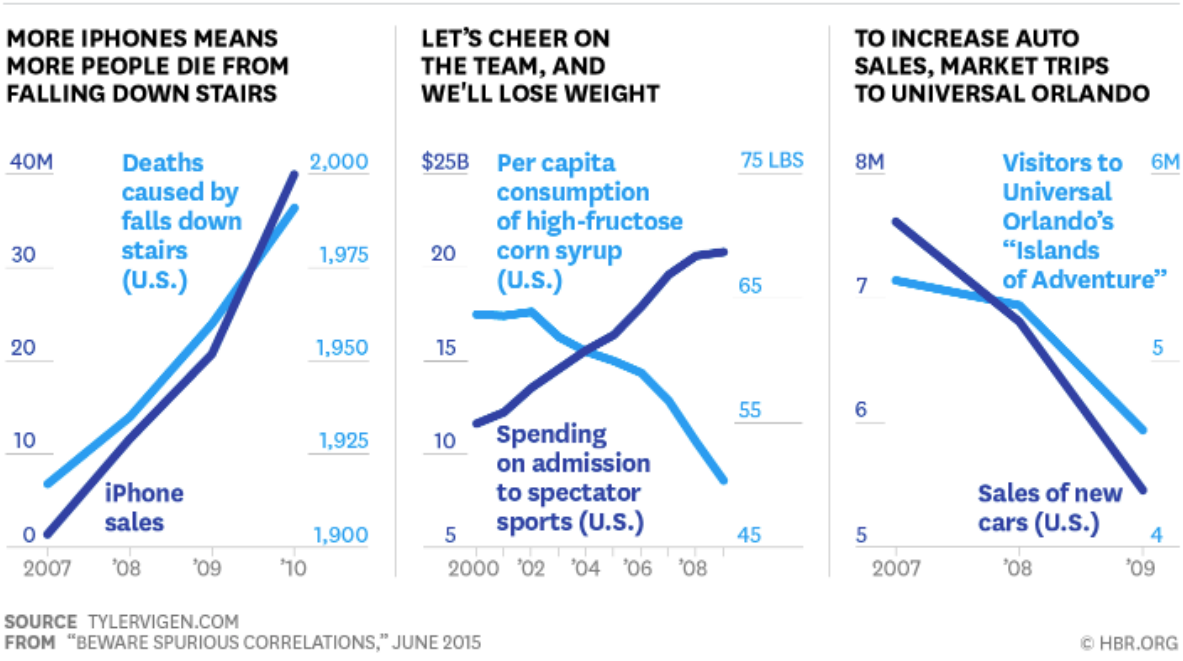
Um estudo realizado no Centro Médico da Universidade da Pensilvânia, publicado na edição de 13 de maio de 1999 da *Nature* (QUINN et al., 1999), concluiu que crianças que dormem com a luz acesa são muito mais propensas a desenvolver miopia na vida adulta. No entanto, em um estudo posterior na Universidade do Estado de Ohio (ZADNIK et al., 2000), também publicado na *Nature*, a equipe de pesquisadores não conseguiu encontrar uma ligação entre crianças em idade escolar que dormiam com a luz acesa e o desenvolvimento de miopia. Neste caso, uma ligação forte entre a miopia parental e o desenvolvimento da miopia das crianças foi identificado (herança genética), observando igualmente que os pais míopes eram mais propensos a deixar uma luz acesa no quarto dos seus filhos. Neste caso, a causa da miopia e de deixar a luz do quarto acesa é de origem parental comprovando que a afirmação do primeiro estudo está incorreta.

Numerosos estudos epidemiológicos demonstraram que mulheres que tomaram a terapia de reposição hormonal (TRH) combinada tiveram uma incidência menor de doença cardíaca coronariana (DCC). Frente a este fato médicos a concluírem que a TRH protegia as pacientes contra DCC. No entanto, estudos randomizados controlados demonstraram que a TRH causou um aumento no risco de DCC, que embora fosse pequeno era estatisticamente significativo. Uma reavaliação sobre os dados dos estudos epidemiológicos mostrou que as mulheres que faziam TRH eram mais propensas a pertencerem a grupos socioeconômicos de níveis mais altos. Consequentemente essas mulheres faziam regimes e exercício melhores do que a população média. O uso de TRH e a diminuição da incidência de doença coronariana foram efeitos coincidentes dos status socioeconômico mais elevados, ao invés de ser uma causa e efeito direto, como se supunha (LAWLOR; SMITH; EBRAHIM, 2004).

Exemplos de má interpretação de causalidades em decorrência da correlação podem ser encontrados no livro "*Beware Spurious Correlations*" (VIGEN, 2015) onde diversos gráfico demonstram esse tipo de correlação como apresentado pela Figura 4. Há uma grande confusão entre pesquisadores, principalmente novatos ou aqueles com pouco conhecimento da

estatística que associam a correlação à causalidade, no entanto, isso não é verdade, a correlação indica que há uma relação entre as variáveis não causalidade.

Figura 4 - Exemplos de correlações espúrias



Fonte: VIGEN, 2015.

Coeficiente de correlação linear de Pearson

O coeficiente de correlação linear de Pearson é um método estatístico que mensura a intensidade, a direção correlação entre duas variáveis aleatórias X e Y de escala métrica (intervalar ou de razão). O coeficiente de correlação linear de Pearson é dado por (eq. 1):

$$r_{xy} = \frac{Cov(X,Y)}{\sqrt{V(X) * V(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X * \sigma_Y} \qquad (1)$$

onde: $\sigma_{X,Y}$ representa a covariância entre as variáveis aleatórias X e Y; σ_X desvio padrão da variável aleatória X e σ_Y o desvio padrão da variável aleatória Y. O coeficiente de correlação de Pearson sempre se mantém no intervalo [-1,1], valores mais próximo de 1 indicam correlação mais fortes. Valores de coeficiente positivos indicam uma relação direta entre as variáveis, ou seja, ambas variam no mesmo sentido. Valores negativos indicam uma correlação inversa, ou seja, enquanto uma variável aumenta o seu valor a outra o diminui Casella e Berger (2001). Supondo-se dados com normalidade bivariada, o teste de hipóteses

para determinar a existência de correlação entre duas variáveis aleatórias *X* e *Y* é dado por (eq. 2):

$$t_c = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1-r_{XY}^2}} \qquad (2)$$

Onde: r_{XY} representa o coeficiente de correlação entre *X* e *Y* e *n* representa o número de elementos na amostra. Neste caso a hipótese nula H_0: ρ = 0 (Não existe relacionamento linear entre as variáveis) não é rejeitada se $|t_c| < t_{(\frac{\alpha}{2};n-2)}$Costa Neto(2011).

Outras correlações bivariadas

O coeficiente de correlação linear de Pearson é um método aplicado quando as variáveis envolvidas no processo são contínuas. No entanto, variáveis não contínuas são comuns, principalmente na área da saúde, foco desta tese. Esse tipo de variáveis também podem ser analisadas por outros tipos de correlações bivariadas (KLINE, 2015):

1. a correlação ponto-bisserial (r_{pb}) é um caso especial de *r* que estima a associação entre uma variável dicotômica e uma contínua (tratamento x controle e peso);
2. o coeficiente de *phi* (φ̂) é um caso especial para duas variáveis dicotômicas (tratamento x controle e sobreviveu x morreu);
3. o coeficiente de correlação de postos de Spearman ou o *rho* (r̂)de Spearman é indicado para duas variáveis classificadas (a ordem de chegada em uma corrida, a classificação por quantidade de tempo de treinamento).

Também é possível analisar correlações não-Pearson que assumem que os dados sejamc ontínuos e normalmente distribuídos em vez de discretos. Por exemplo:

1. a correlação bisserial (r_{bis}) é indicada para avaliar a relação de uma variável contínua e uma variável dicotômica (peso e se recuperou ou não) e estima o que poderia ser o *r* de Pearson se ambas as variáveis fossem contínuas e normalmente distribuídas;
2. a correlação polisserial é a generalização da correlação bisserial e faz basicamente a mesma coisa, no entanto é específica para situações onde se pretende avaliar a

correlação de uma variável contínua e uma variável categórica com três ou mais níveis (peso e graus de recuperação);

3. a correlação tetracórica (r_{tet}) é indicada para calcular a correlação entre duas variáveis dicotômicas, neste caso o método calcula qual seria o valor de *r* de Pearson se ambas as variáveis fossem contínuas e normalmente distribuídas;
4. a correlação policórica é a generalização da correlação tetracórica, que estima o valor de *r* de Pearson porém entre variáveis categóricas ordinais com dois ou mais níveis.

Variância e Covariância

Variância e covariância são conceitos importantes para a compreensão de modelos de AT, portando a seguir se descreve esses conceitos de acordo com Pearl, Glymour e Jewell (2016).

A variância de uma variável *X*, denotada por Var(*X*) ou σ_x^2, é uma medida aproximada de como os valores de *X* em um conjunto de dados ou população estão "espalhados" em torno de sua média. Se os valores de *X* permanecerem perto de um valor, a variação será relativamente pequena. Se eles cobrirem um intervalo grande, a variação será comparativamente grande. Matematicamente, se define a variância de uma variável como a diferença quadrada média dessa variável a partir de sua média. Pode ser calculado primeiro encontrando sua média, μ, e depois calculando (Eq. 3) :

$$Var(X) = E((X - \mu)^2) \qquad (3)$$

O desvio padrão $\sigma_X$de uma variável aleatória *X* é a raiz quadrada de sua variância. Ao contrário da variância, σ_X é expresso nas mesmas unidades que *X*. Por exemplo, a variância da distribuição etária dos eleitores com menos de 45 anos pode ser calculada como sendo:

$$Var(X) = ((23.5 - 31.5)^2 x\ 0.41) + ((37 - 31.5)^2 x\ 0.59)$$
$$= (64\ x\ 0{,}41) + (30.25\ x\ 59)$$
$$= 26{,}24 + 17{,}85 = 43{,}09\ anos^2$$

Enquanto o desvio padrão é:

$$\sigma_x = \sqrt{(43{,}09} = 6{,}56\ anos$$

Isso significa que, a escolha aleatória de um eleitor, tem grandes chances de que este tenha a sua idade com menos 6,56 anos da média de 31,5.

De especial importância a covariância de *X* e *Y*, σ_{xy} mede o grau em que *X* e *Y* variam em conjunto ou estão "associadas". Essa medida de associação realmente reflete uma maneira específica na qual *X* e *Y* covariam. Ela mede até que ponto *X* e *Y* covariam linearmente. De maneira geral, pode se pensar nisso como representar *Y* versus *X* considerando até que ponto uma linha reta captura a maneira como *Y* varia conforme *X* muda.

A covariância σ_{xy} é frequentemente normalizada para produzir o coeficiente de correlação (Eq. 4):

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \qquad (4)$$

Que é um número adimensional que varia de −1 a 1 e representa a inclinação da linha de melhor ajuste depois que se normaliza *X* e *Y* pelos respectivos desvios padrão. ρ_{XY} é um se e somente se uma variável pode predizer a outra de maneira linear, e é zero sempre que uma previsão linear não é melhor do que uma estimativa aleatória.

A próxima seção aborda conceitos sobre modelos gráficos e suas possíveis configurações, uma vez que tanto AT como RB utilizam GADs para representação gráfica de seus modelos causais. Também conceitua a cobertura de Markov que é um método utilizado para inferir causalidades em grafos.

2.3 MODELOS GRÁFICOS

A pesquisa epidemiológica, foco deste estudo, está repleta de incerteza sobre os pressupostos teóricos. Por este motivo esta seção se destina a descrever os modelos gráficos e suas aplicações. A teoria de grafos ou diagramas causais tem sido utilizada há bastante tempo como ferramenta auxiliar para a análise causal. Especialmente a teoria dos grafos acíclicos dirigidos (GAD) que tem sido empregada em conjunto com sistemas especialistas. Nessa área de pesquisa, uma teoria formal permite avaliar os efeitos causais e consequentemente, por meio de um sistema inteligente, se identifica a presença ou não de *links* causais.

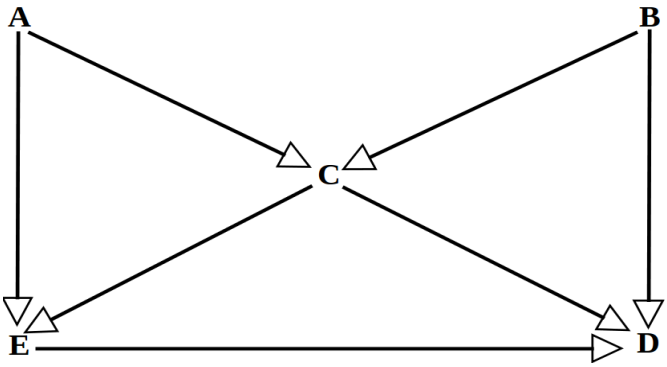
Um diagrama causal se constitui pela abstração dos pressupostos causais incorporados na relação hipotetizada entre as variáveis de um estudo (GREENLAND; PEARL; ROBINS, 1999). Como exemplo se apresenta o estudo sobre a relação da incidência do tratamento anti-histamínico para asma entre crianças do primeiro grau de escolas públicas. Neste estudo,

afirma-se que os *níveis de poluição* e o *sexo* são independentes; que o *sexo* influencia a *administração de anti-histamínicos*, mas de forma indireta, ou seja, por meio de suas relações com a *reatividade brônquica*, porém influencia diretamente o *risco do paciente ter asma*; da mesma maneira a *poluição do ar* influencia indiretamente o *risco do paciente ter asma* por meio da sua influência no uso de *anti-histamínicos* e *reatividade brônquica*. Este estudo pode ser representado pela Figura 5. Nesta Figura *A* representa a poluição, *B* o sexo, *C* a reatividade brônquica, *E* o anti-histamínico e *D* a asma.

Considerando a terminologia de grafos, qualquer linha conectando duas variáveis é chamada de *arco* ou *aresta*. Duas variáveis são consideradas adjacentes se elas estão diretamente conectadas por um arco. Como exemplo as variáveis *A* e *C* na Figura 5 são consideradas adjacentes, porém A e D não são. Um seta com ponta única representa uma ligação direta da causa para o efeito.

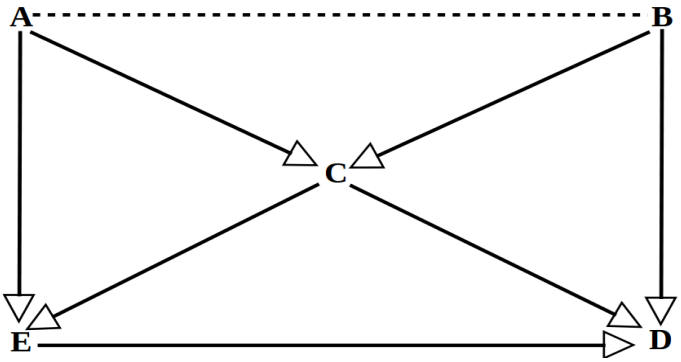
Na Figura 5 por exemplo, a seta partindo da variável *A* para a variável *C* representa o efeito direto da variável *poluição* sobre a variável *reatividade brônquica*. Nesta mesma Figura, a afirmação de que *poluição* afeta asma somente por meio de *reatividade brônquica* e *anti-histamínicos*, corresponde a um efeito indireto. As letras representam os nós ou vértices do grafo e correspondem às variáveis do modelo. Um *path through*, traduzido por "caminho através", é uma rota ininterrupta traçada ao longo ou contra setas conectando nós adjacentes. Por exemplo, o caminho entre as variáveis *E*, *C* e *D* que passam somente por *C* é um *path through.* Um *directed path* ou "caminho direcionado", também conhecido como caminho causal ou *causal path* é aquele que pode ser traçado pela sequência que se inicia pela entrada em uma seta por meio de sua base e saindo pela sua ponta. Na Figura 5 o caminho *A-C-D* é considerado um *directed path*, mas *E-C-D* não é. Um nó dentro de um caminho é classificado como interceptador do caminho, na Figura 5 o nó C intercepta os caminhos *A-C-D* e *E-C-D.* Uma variável *X* é um antecessor ou causa de outra variável *Y* se houver um caminho direcionado de setas que saem de *X* para *Y*, neste caso, *Y* é dito ser um descendente de *X* ou afetado por *X*. Na Figura 5, *A*, *B* e *C* são antecessores de *E* e *D*, que consequentemente são descendentes de *A*, *B* e *C*. Uma variável *X* é considerada pai de *Y* em um grafo se existir um seta de *X* para *Y*, paralelamente *Y* é dito ser filho de *X* ou diretamente afetado por *X*. Na Figura 5 as variáveis *A* e *C* são pais de *E*, *C* e *E* são filhos de *A.* Um arco não direcional (sem pontas) é utilizado para indicar tha duas variáveis são associadas por outras razões além do relacionamento antecessor x descendente. Por exemplo, na Figura 5 utilizou-se essa notação para representar a relação cuja fonte não é especificada pelo grafo (GREENLAND; PEARL; ROBINS, 1999).

Figura 5 - Grafo representando a causalidade do tratamento anti-histamínico para asma em crianças do primeiro grau de escolas públicas



Fonte: GREENLAND; PEARL; ROBINS, 1999.

Figura 6 - Associação desconhecida, mas existente, entre asma e sexo representado pela linha pontilhada



Fonte: GREENLAND; PEARL; ROBINS, 1999.

Um caminho que conecta *X* com *Y* é considerado um "caminho de portas do fundo" (*backdoor path*) de *X* para *Y* se ele tiver uma seta apontando para *X* Pearl(1995). Por exemplo na Figura 5, todos os caminhos de *E*para *D*, exceto o caminho direto são "caminho de portas do fundo". Um caminho colide na variável *X* se este caminho entra e sai desta variável por meio das pontas de uma seta, neste caso, *X* é conhecido como colisor de caminho (*collider on the path*) Spirtes, Glymour e Scheines(1993). Um caminho é bloqueado se este conter um ou mais colisores, caso contrário, é um caminho desbloqueado (GREENLAND; PEARL; ROBINS, 1999). O "caminho de portas do fundo" *E-A-C-B-D* na Figura 5 é bloqueado porque colide em *C*, que é o único colisor no caminho. Por outro lado, o "caminho de portas do fundo" *E-A-C-D* é considerado desbloqueado porque nem A nem C são colisores neste caminho. Qualquer tipo de caminho pode incluir arcos não-direcionados, como por exemplo, o "caminho de portas do fundo" *E-A-B-D* na Figura 6 que é um caminho desbloqueado.

Neste estudo serão utilizados somente grafos acíclicos dirigidos (GADs ou DAG, do inglês *Directed Acyclic Graph*). Este é acíclico porque nenhum caminho direto do grafo forma um circuito fechado (*closed loop*). É considerado dirigido porque todos os arcos entre

as variáveis são representados por setas (de ponta única ou dupla) que apontam em uma direção (GREENLAND; PEARL; ROBINS, 1999).

Os pressupostos da teoria dos grafos são qualitativos e não paramétricos, ou seja, não implicam nada sobre a forma das relações ou distribuições das variáveis, que podem ser discretas ou contínuas. A criação de um efeito por uma causa exige um caminho causal, o qual é representado por um caminho dirigido em um grafo, da causa para o efeito, dessa forma a ausência de um caminho dirigido de *X* para *Y* representa a suposição de que não existe efeito de *X* em *Y*. Por exemplo, na Figura 5 observa-se que não há arco direcionado de *A* para *B*, *B* para *A*, *A* para *D* e *B* para *E* (GREENLAND; PEARL; ROBINS, 1999). Dessa forma, a Figura 5 representa os pressupostos causais básicos de que *A* não afeta diretamente *B*, *B* não afeta diretamente *A*. *A* não afeta diretamente *D* e *B* não afeta diretamente *E*. A primeira suposição também implica que não há efeito de *A* em D que é transferido através de *B* e assim por diante.

A próxima seção descreve as possíveis configurações entre os nós e arcos em um grafo, pois as formas como as variáveis se conectam é importante por determinar como podem ser feitas as inferências necessárias para avaliar a causalidade.

2.3.1 Possíveis configurações entre nós e arcos

As RBs por exemplo permitem que se crie um modelo causal, cujo objetivo é representar as relações entre as variáveis desse modelo. A forma como as variáveis se conectam determina as possíveis configurações entre nós e arcos, gerando a estrutura da rede, também conhecidas como *conexões fundamentais* (FIGURA 7) e formam os blocos de construção das propriedades gráficas e probabilísticas da RB (SCUTARI, 2014). Neste contexto torna-se importante estudar os possíveis tipos de conexões que podem existir entre os nós das RBs. Essas conexões revelam informações sobre a natureza das variáveis, bem como a sua dependência com relação a outras variáveis, é importante analisar quando duas variáveis são dependentes ou independentes (DE SÁ, 2014). Variáveis são dependentes quando a observação de uma influencia a outra, neste caso deve existir uma aresta direcionada entre elas. Deve-se observar também o fluxo dessa dependência, pois hora este é direto e hora é indireto, ou seja, a dependência entre duas variáveis depende de uma terceira variável. Já independência condicional (termo comum usado em RBs) ocorre quando o valor de uma variável não influencia o resultado da outra. Essas estruturas, que são essenciais para

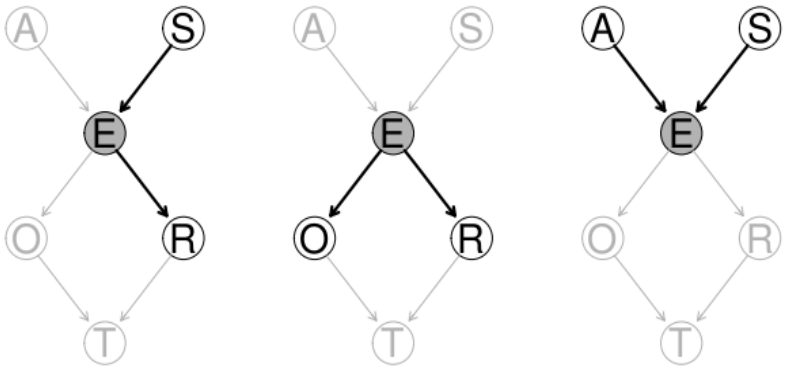
caracterização e aprendizagem de RBs, são classificadas em três formações distintas, descritas nos parágrafos seguintes (SCUTARI, 2014).

Conexões seriais contém estruturas onde o fluxo de influência causal é do tipo $S \rightarrow E \rightarrow R$ (primeiro exemplo da FIGURA 7). Neste caso, ambos os arcos têm a mesma direção e seguem um após o outro e qualquer tipo de influência que *S* tiver será replicada para *E* e consequentemente para *R*. No entanto, se o valor de *E* for informado, o fluxo causal entre *S* e *R* é interrompido e neste caso as variáveis serão consideradas como d-separadas. Neste caso, diz-se que *S* e *R* são d-separadas por *E* ou condicionalmente independentes dado *E*.

Conexões divergentes possuem a estrutura do fluxo de influência causal do tipo do tipo $R \leftarrow E \rightarrow O$ (segundo exemplo da FIGURA 7). Nesta situação, os dois arcos possuem direções divergentes a partir de um nó central, ou seja, as variáveis *O* e *R* divergem da variável *E*. Considerando esta situação, a variável *E*, como nó pai, transmitirá a influencia causal para todos os nós filhos *O* e *R*, exceto quando o estado de *E* for conhecido, pois similar ao caso de conexões seriais, *O* e *R* serão d-separadas por *E* ou condicionalmente independentes dado *E*.

Conexões convergentes são representadas por estruturas do tipo $A \rightarrow E \leftarrow S$ (terceiro exemplo da FIGURA 7). Nessa estrutura, os arcos convergem para um nó central, ou seja, as variáveis *A* e *S* convergem para *E* indicando que *E* sofrerá influência causal de *A* e *S*. Neste caso, *A* e *S* não são d-separados quando o estado de *E* é conhecido, consequentemente *A* e *S* são condicionalmente dependentes de *E*.

Figura 7 - Exemplos de conexões fundamentais

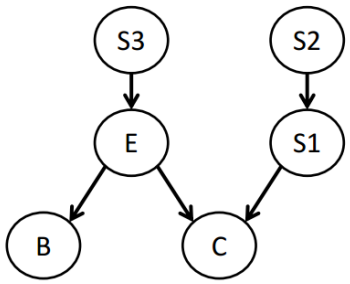


Fonte: SCUTARI, 2014.

Como exemplo geral desses três conceitos Pearl(2003) considere que, se *X*, *Y*, *Z*são três conjuntos disjuntos de nós de uma RB, então *Y* é dito d-separador de *X* de *Z*, se e somente

se *Y* bloqueia todos os caminhos de um nó em *X* para um nó em *Z*. De acordo com essa afirmação e a Figura 8, o nó de *E* d-separa os nós *B* e *C* do nó *S3* (caminhos divergentes) e o nó *S1* d-separa o nó *C* do nó *S2* (caminhos seriais). Já o nó *C* não d-separa o nó *E* do nó *S1* porque os caminhos convergentes não estão bloqueados considerando-se o nó no ponto de convergência ou seus descendentes. Todas as relações de d-separação entre nós em um grafo implicam relações de independência condicional entre as variáveis correspondentes.

Figura 8 - RB representando a relação entre a incidência de câncer (C), a exposição ambiental (E), um biomarcador (B) e três nucleotídeos (S1, S2, S3)



Fonte: SU, 2013.

O processo de avaliar se uma variável é d-separada de outra, comumente usado na construção de RBs, pode ser complexo em modelos que representam o mundo real. Outras soluções como a cobertura de Markov também são exploradas.

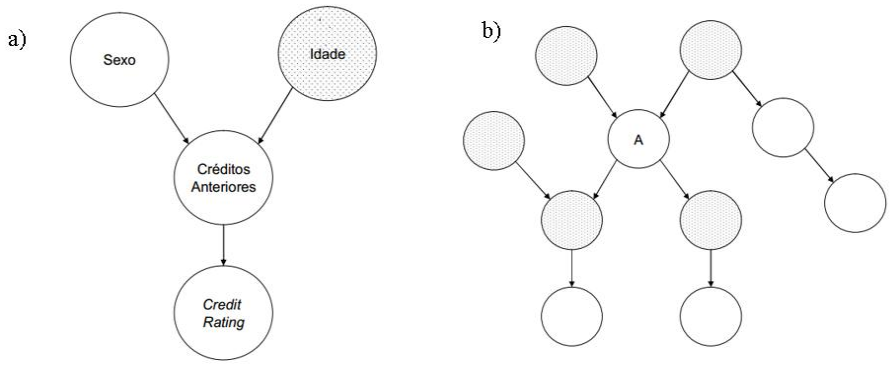
Apróxima seção aborda o conceito sobre a cobertura de Markov, conceito també utilizado por métodos de inferência causal para inferência causal entre nós de um grafo.

2.3.2 Cobertura de Markov

A cobertura de Markov (CM) de um nó em uma RB envolve as variáveis-pai, variáveis-filhos e pais dos filhos de uma determinada variável. Neste caso, envolve todas as variáveis que podem dar informações sobre a variável que representa o nó. Pode-se afirmar que a CM ocorre quando for fornecido informações sobre seus pais, filhos e pais dos filhos de um nó, neste caso este nó é independente de todos os outros nós. Como exemplo, na Figura 9a, a cobertura de Markov para a variável *Idade* envolve a variável *Créditos Anteriores* que é variável-filha da variável *Idade* e a variável *Sexo* que é a variável-pai de uma variável-filho da variável *Idade*. Observa-se que a variável *Idade* não possui variáveis-pai, mas se elas existissem seriam consideradas na cobertura de Markov também. Um segundo exemplo é

apresentado na Figura 9b onde as variáveis da cobertura de Markov para a variável *A* são apresentadas em cinza (KARCHER, 2009).

Figura 9 - Exemplos de cobertura de Markov



Fonte: KARCHER, 2009.

Uma variação da CM é conhecida como cobertura aproximada de Markov (CAM), neste caso, considera-se um nó independente de todos os outros nós dados apenas seus pais e filhos Dos Santos(2011). A vantagem de se utilizar a CAM ou CM é que esses métodos diminuem o custo computacional diminuindo a quantidade de variáveis a serem exploradas durante a execução do algoritmo.

Na seção seguinte se apresentam diversos conceitos sobre redes Bayesianas que se consideram importantes para a compreensão deste estudo.

2.4 REDES BAYESIANAS

Com o passar do tempo, o ser humano aprendeu que a observação de padrões poderia auxiliá-lo a testar e confirmar hipóteses. Consequentemente com a evolução das técnicas e equipamentos, criou-se a inteligência artificial (IA) e dentro desta área surgiu a mineração de dados (MD) cujo objetivo é evidenciar padrões e auxiliar na descoberta de conhecimento. Experimentos envolvendo IA e MD levaram pesquisadores concluírem que essas técnicas deveriam ter a capacidade de argumentar logicamente e raciocinar probabilisticamente ao ser utilizada para resolução de problemas do mundo real e consequentemente ter habilidade para lidar com a incerteza (KORB; NICHOLSON, 2003). Essa incerteza é representada por situações onde as evidências são incompletas, levam a conclusões falíveis e a necessidade constante de ser capaz de se recuperar ou diminuir ao máximo a chance de erro. Neste contexto, o desenvolvimento de uma arquitetura de *software* para IA envolvendo a inferência

Bayesiana, como as RBs se tornou um componente importante como solução para modelar problemas reais do cotidiano (KORB; NICHOLSON, 2003).

Dessa forma, nas próximas seções são apresentados alguns dos conceitos fundamentais que caracterizam a modelagem causal por meio de métodos para construção de RBs.

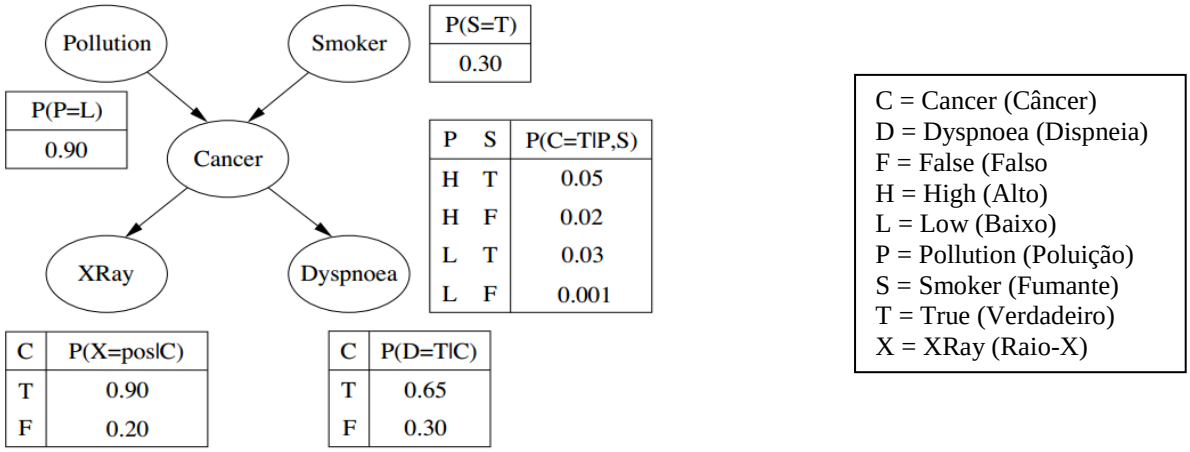
2.4.1 Definição de Redes Bayesianas

O conceito de RB também é conhecida pelos nomes de modelos gráficos recursivos (*recursive graphical models*), redes Bayesianas de crença (*Bayesian belief networks*), redes de crença (*belief networks*), redes probabilísticas causais (*causal probabilistic networks*), redes causais (*causal networks*), diagramas de influência (*influence diagrams*), redes Bayesianas dinâmicas (*dynamic Bayesian networks*) (DALY; SHEN; AITKEN,2011). Uma RB é um modelo gráfico capaz de identificar relações probabilísticas entre um conjunto de variáveis. Essas relações representam o conhecimento extraído de um conjunto de dados (KORB; NICHOLSON, 2003). RBs permitem executar inferência probabilística sobre essas variáveis considerando condições de incerteza (CHENG et al., 2002). Embora seja uma publicação de mais de 20 anos atrás, em seu artigo Heckerman (1997) aponta quatro principais vantagens, ainda presentes nos dias atuais, oferecidas pelas técnicas de modelagem de dados por meio de RBs. Primeiro, RBs conseguem lidar bem com dados incompletos por ter a capacidade de identificar dependências entre as variáveis por meio de testes probabilísticos. Segundo, RBs permitem que se aprenda sobre as relações causais, um processo útil quando se necessita entender sobre um domínio problemático, durante a análise exploratória de dados ou quando se deseja fazer previsões baseadas em intervenções. Terceiro, as RBs, por sua característica de possuir uma semântica de causalidade e ser baseada na teoria probabilística, contribuem com a modelagem de eventos do mundo real por meio da combinação do conhecimento *a priori* (de fundamental importância para a modelagem de dados) e dados do domínio. Quarto, as RBs em conjunto com métodos Bayesianos e outros tipos de modelos oferecem uma solução eficiente para evitar o super ajuste (*overfitting*) de dados.

Basicamente RBs (FIGURA 10) são compostas por dois componentes (KORB; NICHOLSON, 2003). O primeiro é a estrutura gráfica, que representa um domínio incerto. Essa estrutura é composta por nós que representam um conjunto de variáveis aleatórias do domínio e arcos direcionados que conectam pares de nós e representam as dependências diretas entre variáveis. O segundo é um conjunto de parâmetros representado por uma tabela

de probabilidades condicionais (TPC) para cada variável. A estrutura gráfica tem como restrição única que seus arcos não devem permitir ciclos direcionados, consequentemente as RBs devem ser representadas graficamente por GADs.

Figura 10- RB representando o problema de câncer de pulmão com suas TPCs



Fonte: KORB; NICHOLSON, 2010.

A estrutura gráfica das RBs pode ser representada por $G = (V, A)$ onde V é o conjunto de nós (ou vértices). O GAD representando essa estrutura define uma fatoração da probabilidade conjunta de distribuição de $V = \{X_1, X_2, ..., X_V\}$, chamada de distribuição de probabilidade global, em um conjunto de distribuições de probabilidades locais, um para cada variável (SCUTARI, 2009). A maneira de executar a fatoração (equações 5 e 6) é fornecida por Korb eNicholson (2003) por meio da propriedade Markov de RBs, cuja afirmação confirma que cada variável aleatória X_i depende diretamente apenas de seus pais. Deve-se observar que há equações diferentes para cada tipo de variável, pois os nós da RB podem representar variáveis discretas ou contínuas. Assumindo-se variáveis discretas as probabilidades condicionais (força da relação entre as variáveis) serão representadas por TPCs. Em caso de variáveis contínuas essa representação se dá por meio de funções de densidade ou distribuição de probabilidade condicional.

$$P(X_{1,\dots,}X_{v,})\prod_i^v P(X_I \vee \prod X_i) \qquad \text{para variáveis discretas} \qquad (5)$$

$$F\left(X_{1,\dots,}X_{v,}\right)\prod_i^v F(X_I \vee \prod X_i) \qquad \text{para variáveis contínuas} \qquad (6)$$

A próxima seção explica como se deve raciocinar utilizando o pensamento bayesiano que e baseia em probabilidades condicionais.

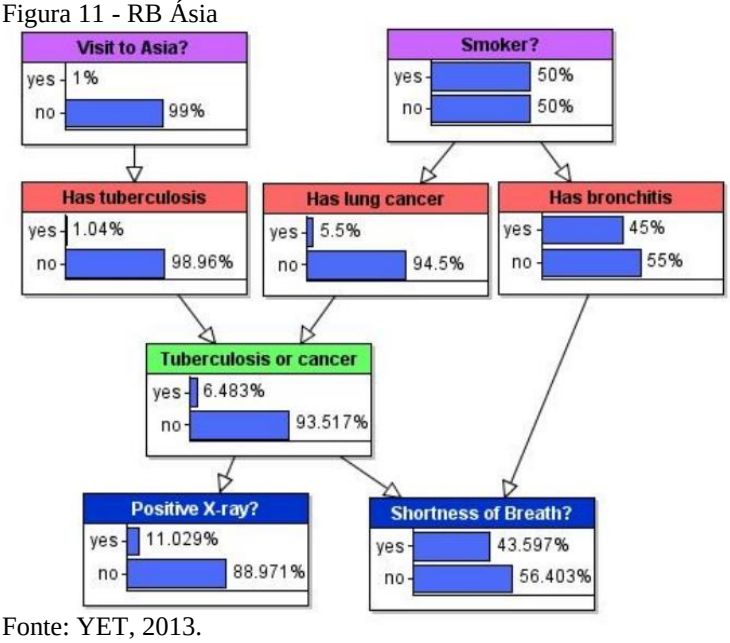
2.4.2 Raciocinando com Redes Bayesianas

Para melhor entendimento de como se raciocina de acordo com o pensamento bayesiano, (YET, 2013) fornece um exemplo: ultimamente o Sr. John Doe tem sofrido com falta de ar; como consequência disso, ele não para de se preocupar com a possibilidade de ter câncer, mesmo pensando em outras causas para sua falta de ar, como bronquite, por exemplo. Devido a sua preocupação ele resolve procurar um médico que decide usar a RB Ásia (FIGURA 11) como ferramenta de apoio à decisão para chegar a um diagnóstico sobre a doença do Sr. Doe. Inicialmente o médico considera 3 hipóteses: câncer, tuberculose e bronquite. O modelo de RB tem uma variável representando cada uma das hipóteses (“Tem tuberculose”, “Tem câncer”, “Tem bronquite”), e pode fazer cálculos probabilísticos sobre a hipótese com base na informação que é inserida no modelo.

Primeiro o clínico pergunta sobre os sintomas do Sr. Doe e recalcula as probabilidades quando insere dados sobre falta de ar. Neste momento a bronquite é o diagnóstico mais provável (FIGURA 12a). Para evitar um diagnóstico errado caso o paciente tenha uma doença mais grave do que bronquite, como tuberculose ou câncer, o clínico solicita um raio-X de tórax. O resultado do raio-X, por ser positivo, faz com que o clínico fique ainda mais preocupado com a possibilidade do seu paciente ter câncer (FIGURA12b).

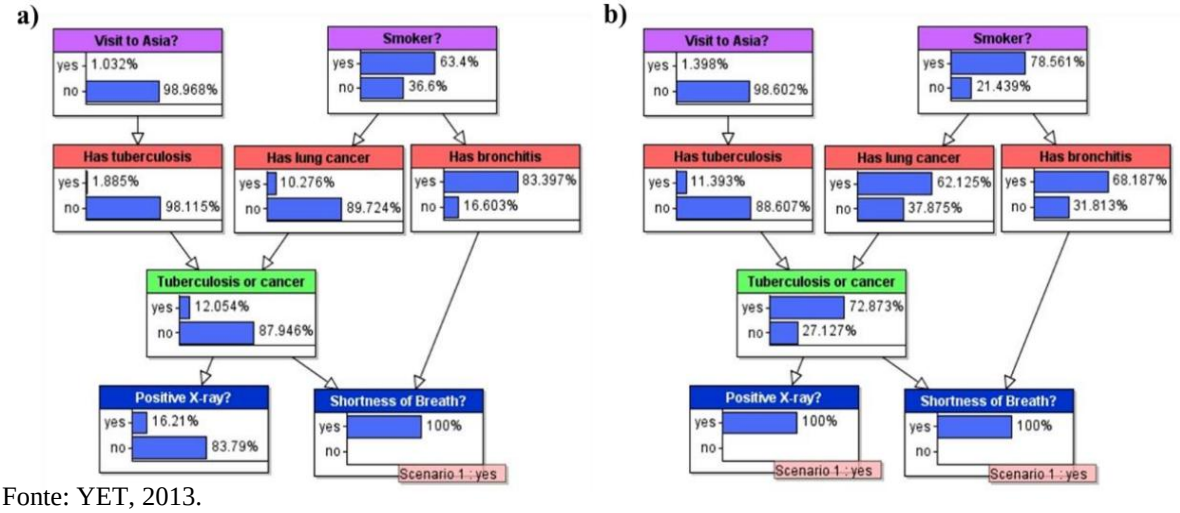
A fim de coletar mais informações o médico pergunta ao Sr. Doe sobre casos de câncer na família e seus hábitos de fumante. O Sr. Doe diz que não fuma regularmente, mas que andou fumando ao passar um feriado no Camboja. A informação sobre a viagem ao Camboja pode ser uma importante fonte de informação para fortalecer a hipótese de tuberculose, uma vez que essa doença é mais prevalente neste país de acordo com relatório da Organização Mundial de Saúde de 2012. Após inserir a informação sobre hábitos de fumante e a visita à Ásia a probabilidade de câncer, antes maior, deu lugar a probabilidade de o Sr. Doe ter tuberculose (FIGURA 13a e FIGURA 13b).

Figura 11 - RB Ásia



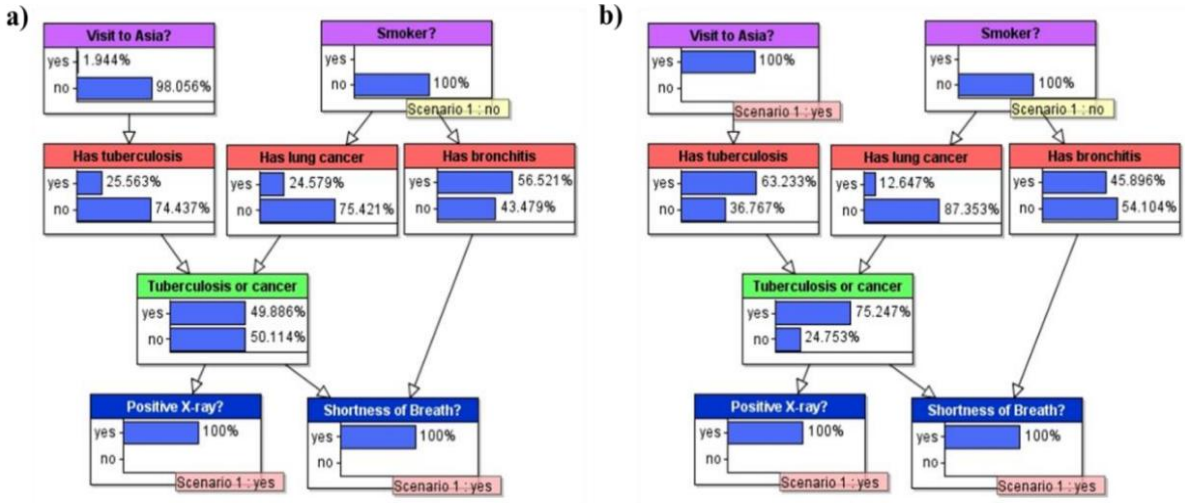
Fonte: YET, 2013.

Figura 12 - Probabilidades atualizadas após observar a) sintomas de falta de ar e b) raio X.



Fonte: YET, 2013.

Figura 13 - Probabilidades atualizadas após obter informações sobre: a) histórico de fumante e b) visita à Ásia



Fonte: YET, 2013.

Este exemplo ilustra três maneiras com as quais as RBs propagam a informação para atualizar as probabilidades (YET, 2013):

1. Raciocínio causal: Quando se insere uma informação em um "nó causa" a probabilidade dos seus "nós efeito" serão atualizadas. Foi o caso da visita para Ásia ter aumentado a probabilidade de tuberculose;
2. Raciocínio diagnóstico: Após inserir uma informação em um "nó efeito" a probabilidade dos seus "nós causa" será atualizada. Ao inserir a informação sobre a falta de ar do paciente aumentou a probabilidade de bronquite;
3. *Explaining Away (sem tradução)*: Se qualquer um dos "nós efeito" ou seus descendentes é observado inserindo uma informação em um "nó causa" este atualizará a probabilidade de outros "nós causa". Por exemplo, ao saber o resultado do raio-X, a presença da falta de ar e a visita à Ásia consequentemente aumentará a probabilidade de tuberculose e diminuirá a probabilidade de câncer, que é causa de raio-X positivo e falta de ar. Em resumo, a probabilidade maior de tuberculose, resultado da visita para a Ásia explicou outras causas geradas pelo raio-x positivo e pela falta de ar.

A seção que segue descreve o propósito das RBs e detalha a forma como se aprende a estrutura a partir de dados, método que será utilizado por este estudo.

2.4.3 O Propósito das redes Bayesianas

Estudos envolvendo RBs normalmente envolvem três áreas de estudo (DALY; SHEN; AITKEN, 2011): a) aprendizado de estrutura, que consiste gerar várias estruturas gráficas ou modelos probabilísticos a partir de um conjunto de dados e escolher a que melhor explica esse conjunto de dados; b) aprendizado de parâmetros, após o aprendizado da estrutura tabelas de probabilidades condicionais para cada nó serão criadas Na literatura encontram-se estudos que tratam esses temas de forma isolada ou conjunto e c) Inferência probabilística, cuja função é determinar os valores de um conjunto de variáveis que melhor explica os valores de outro conjunto de variáveis.

Aprendizagem de estruturas de redes Bayesianas

A aprendizagem de uma rede bayesiana pode ocorrer com base em: a) assimilar a estrutura gráfica da rede e b) perceber os parâmetros da rede, que significa descobrir as distribuições de probabilidades entre os nós. Ambos podem ser executados como aprendizagem não-supervisionada, usando informação fornecida por um conjunto de dados ou aprendizagem supervisionada com ajuda de um especialista no domínio a ser modelado. Existem estratégias que combinam as duas abordagens, uma vez que nem sempre o especialista tem informação suficiente para construir a RB, principalmente em domínios onde um grande número de variáveis está envolvido, como análise de redes genéticas (SCUTARI, 2014).

A formalização para aprendizagem de RBs apresentada por Scutari (2014), pode ser definida como: considere um conjunto de dados D e uma RB $B = (G, X)$. Se representarmos os parâmetros da distribuição local de X como θ, podemos assumir sem perda de generalidade que θ exclusivamente identifica X na família paramétrica de distribuições escolhidas para modelagem de D e escrever $B = (G, \theta)$. Então a aprendizagem de uma RB pode ser formalizada como (Eq. 7):

$$\underbrace{\Pr(B|D)}_{Aprendizagem} = \underbrace{\Pr(G|D)}_{Aprendizagem\ Estrutura} \cdot \underbrace{\Pr(\theta|G,D)}_{Aprendizagem\ Parâmetros} \qquad (7)$$

Algoritmos de aprendizagem de estrutura podem ser classificados em três abordagens: baseado em restrições, baseado em pontuação ou híbrido, as quais trabalham sob o seguinte conjunto de suposições (SCUTARI, 2014):

1. deve haver uma correspondência um-para-um entre os nós do GAD e as variáveis aleatórias em X, o que significa que não deve haver múltiplos nós que são funções determinísticas de uma simples variável;
2. todos os relacionamentos entre as variáveis em X devem ter independência condicional, porque eles são por definição o único tipo de relacionamento que pode ser expressado por uma RB;
3. toda combinação de possíveis valores das variáveis em X precisam representar um evento observável e válido. Essa suposição implica em uma distribuição global

estritamente positiva, que é necessária para determinar unicamente a cobertura de Markov e, por conseguinte, um modelo unicamente identificável;

4. observações são tratadas como realizações independentes de um conjunto de nós. Se alguma forma de dependência temporal ou espacial está presente, ela precisa ser especificamente contabilizada para a definição da rede.

Algoritmos baseados em restrições

Esta seção descreve como funcionam os algoritmos baseados em restrição para aprender a estrutura de RB. Este tipo de algoritmo será utilizado para criar a estrutura de grafo para o modelo de entrada do modelo de AT.

Os algoritmos dessa categoria se concentram na identificação de relações de independência condicional entre variáveis por meio dos dados observados. Para isso, esse tipo de algoritmo utiliza algum teste de independência condicional, os quais são utilizadas para avaliar a relação existente entre as variáveis para então restringir a estrutura da RB (SU, 2013). Entre os diversos algoritmos existentes na literatura, alguns serão descritos para que se entenda como funcionam, quais seus requisitos, problemas e qual a sua complexidade.

O algoritmo SRA (*Search-space Reduction Algorithm*), que requer ordenação causal das variáveis, tem o objetivo de criar uma RB que revele o máximo de informações sobre a independência condicional e como resultado este gera uma RB muito esparsa Srinivas (1990). Este vai criando a estrutura da rede sempre com o objetivo de manter o menor número possível de arcos direcionados em cada etapa de execução do algoritmo. A complexidade deste algoritmo é $O(2^n)$ onde n é quantidade de nós.

Os algoritmos SGS (Spirtes, Glymour, and Scheines) e PC (Peter and Clark) Spirtes e Glymour(1993) não exigem a ordenação causal das variáveis. SGS surgiu primeiro, porém seu desempenho é considerado pouco eficiente, uma vez que cada par de variáveis exige testes com todas as outras variáveis gerando um processamento exponencial conforme o algoritmo evolui. Já o algoritmo PC, uma variante do SGS, tem maior velocidade de processamento, mas por verificar a d-separação (ver seção 2.3.1) somente entre X e Y somente com base em seus vizinhos este pode gerar erros no processo de remoção de arcos.

O algoritmo de Cheng, Bell e Liu (1997a), chamado de TPDA-π(*Three-Phase Dependency Analysis Algorithm*) avalia as dependências condicionais entre as variáveis por meio de testes de informação mútua condicional. Este algoritmo depende da ordenação causal entre as variáveis e sua execução se dá em três etapas. A primeira, conhecida como ‘esboço’

cria um esqueleto da rede com base no cálculo de uma medida de proximidade entre as cada par de vértices. A segunda, chamada ‘expansão’ (*thickening*), cria um mapa de independência do modelo de independência conhecido como *I-Map* adicionando novas arestas sempre que os pares de variáveis não puderem ser d-separados. A terceira e última etapa, nomeada de ‘refinamento’ (*thining*) é responsável pela criação do *I-Map* Mínimo, nesta etapa se avalia cada aresta do *I-Map* por meio de testes de independência condicional e se remove caso os dois nós possam ser d-separado (ver seção 2.3.1).

Um outro algoritmo de Cheng, Bell e Liu (1997b), o TPDA realiza as mesmas três fases descritas anteriormente, porém não exige ordenação causal das variáveis. Por esse motivo, este gera a dificuldade de avaliar se dois nós são condicionalmente independentes e gerar uma orientação entre as arestas da estrutura gráfica apreendida, criando a necessidade de se adotar métodos adicionais para resolver este problema. Por este motivo sua complexidade se torna $O(n^4)$.

Duas versões simplificada dos algoritmos de Cheng et al. (1997a e 1997b) geraram os algoritmos SLA-π (*Simple Learning Algorithm*) que exige ordenação causal das variáveis e o SLA, que não exige, são ambos compostos apenas pelas fases de expansão e refinamento e possuem a mesma complexidade dos algoritmos de origem.

Segundo (SCUTARI, 2014) os algoritmos dessa categoria de métodos baseado em restrição se baseiam no trabalho de (VERMA; PEARL, 1990) em mapas e sua aplicação em modelos de gráficos causais. Seu algoritmo *Inductive Causation* (IC) fornece um método para aprendizagem da estrutura do GAD de RBs usando testes de independência condicional. De acordo com (SU, 2013) procedimentos de teste de hipóteses, tais como o teste qui-quadrado, são utilizados para remover arestas de um grafo não direcionado totalmente conectado com base nos valores de independência incondicional encontrados. Em seguida, as arestas recebem direções entre os nós de acordo com o critério de d-separação identificado. A diferença entre as dependências probabilísticas identificadas pelos caminhos seriais, divergentes e convergentes, é essencial para inferir a direção da aresta a partir da análise de dados. Todas as relações de d-separação entre nós de um grafo implicam em relações de independência condicional entre as variáveis correspondentes.

O algoritmo IC apresenta um problema que o impede de ser executado para qualquer problema do mundo real devido ao possível número exponencial de relacionamentos de independência condicional, o que levou ao desenvolvimento de mais alguns algoritmos baseados em restrição (SCUTARI, 2014):

1. PC: a primeira aplicação prática do algoritmo IC (SPIRTES;GLYMOUR; SCHEINES, 2000);
2. *Grow-Shrink* (GS): se baseia no algoritmo Grow-Shrink Markov Blanket de (MARGARITIS, 2003), uma abordagem simples para detecção da cobertura de Markov;
3. *Incremental Association* (IAM): derivado do algoritmo *Incremental Association Markov Blanket* de (TSAMARDINOS, 2003) é um esquema de seleção de duas fases;
4. *Fast Incremental Association* (Fast-IAMB): uma variação do IAMB que utiliza etapas especulativas de seleção para reduzir o número de testes de independência condicional (YARAMAKALA;MARGARITIS, 2003);
5. *Interleaved Incremental Association* (Inter-IAMB): mais uma variação do IAMB que utiliza etapas especulativas de seleção para evitar falsos positivos na fase de detecção da cobertura de Markov (TSAMARDINOS, 2003).

Algoritmos baseados em busca e pontuação

Esta seção descreve como funcionam os algoritmos baseados em pontuação para aprender a estrutura de RB. Este tipo de algoritmo tamém será utilizado para criar a estrutura de grafo para o modelo de entrada do modelo de AT.

Esse tipo de algoritmo tem dois principais componentes, o método de busca e o método de pontuação, e se baseiam em buscas heurísticas para o problema de aprendizagem de estrutura de uma RB (SU, 2013). Por meio dessas buscas no espaço de estrutura e se adicionam novas arestas à estrutura da RB. Durante esse processo, cada candidato da RB recebe uma pontuação (*network score*) que reflete a sua qualidade de encaixe. Em seguida o algoritmo tenta maximizar essa pontuação, objetivando alcançar a estrutura de melhor qualidade, ou seja, cuja pontuação seja maior (SCUTARI, 2014). O número de estruturas possíveis se torna superexponencial conforme o número de nós aumenta, o que gera um número muito grande de possíveis estruturas e resulta em uma pesquisa exaustiva, mesmo para estruturas com poucas variáveis (SU, 2013).

O primeiro estudo que inspirou a criação de diversas técnicas de busca e pontuação foi apresentado por Chow e Liu (1968). O estudo desses pesquisadores se baseava em um algoritmo que aprendia a estrutura de RBs a partir de dados e gerava um grafo no formato de árvores. Esse algoritmo não exigia ordenação causal das variáveis, utilizava busca gananciosa

(*greedy search*) para adicionar arcos à RB e pontuava por meio de entropia da rede. Uma das grandes vantagens deste algoritmo era que o mesmo apresentava complexidade limitada a $O(n^2)$ cálculos de dependência em pares. Outra vantagem era que este método conseguia apresentar um estimador de verossimilhança máximo a partir da distribuição de dependência em árvore.

Esse método foi estendido para o formato de poli-árvores por Rebane e Pearl (2013) onde se conseguiu criar um modelo de dependência com mais arestas direcionadas e portanto mais informativo. Outro método baseado em entropia para gerar o score de cada aresta da rede foi o Kulató Herskovits e Cooper (2013). Este método recebe uma base de dados e suas variáveis em ordenação causal e assume inicialmente que todas as variáveis são independentes. Em seguida, utilizando uma busca gananciosa inicia a adição de arestas direcionadas entre os pares de nós com o objetivo de manter a aciclicidade e diminuir ao máximo a entropia geralda RB. O algoritmo para sempre que não existir um nó pai (na lista de ordenação causal) do nó atual que seja capaz de diminuir a sua entropia. A complexidade deste algoritmo é $O(n^4 \text{x} 2^n)$.

Uma vertente do Kulató chamada K2 foi desenvolvida pelos mesmos autores Herskovits e Cooper (2013) porém se diferenciando por sua métrica de escore, a qual usava um escore Bayesiano cujo objetivo era maximizar a probabilidade da estrutura da RB. Essa vertente se estendeu para o K2 reverso (K2R), onde se inicia a estrutura da RB totalmente conectada e com ordenação causal e remove os nós gradualmente por meio de método guloso e de acordo com a métrica Bayesiana. O K2 apresenta complexidade de $O(mxn^4x^r)$, onde m é o número de registros da base de dados, e r é o número máximo de ocorrências para cada variável.

Melhorias foram implementadas no K2 gerando o K2+, Peng e Ding(2003). Entre os benefícios dessa nova versão foi criada uma busca linear para gerar o grafo candidato, criou-se uma forma de remover ciclos e diminuir a perda da verossimilhança e por fim, desenvolveu-se um método para melhorar a estabilidade da rede refinando a sua estrutura. A complexidade do K2+ é $O(n^2)$.

O princípio da descrição mínima *(minimal description length*) ou MDL foi o método utilizado por Suzuki (1999) para criação de seu algoritmo de aprendizado de estrutura de RBs. Esse algoritmo exige a ordenação causal de variáveis, utiliza a técnica de '*branch and bound*' e atua selecionando as dependências entre as variáveis de acordo com o princípio de selecionar a estrutura de rede mais simples mas com melhor ajuste aos dados.

Para Su (2013) existe uma série de possíveis critérios para usar pontuação de estruturas de RBs, porém como na maioria das vezes a estrutura e os parâmetros da RB são desconhecidos, a probabilidade marginal completa deve ser calculada. Esse cálculo completo quando necessário para o espaço de parâmetros e espaço estrutura é impraticável, exceto para redes menores, o que exige que sejam utilizados métodos, como *Bayesian Information Criterion* (BIC), que é formulada como (Eq.8):

$$BIC = \log\left(\rho\left(D\middle|\hat{\theta}, G\right)\right) - \frac{n_p}{2}\log(N) \qquad (8)$$

Onde $p\left(D\middle|\hat{\theta}, G\right)$ é a probabilidade do dado *D* de acordo com o parâmetro $\widehat{\theta}$ e a estrutura *G*. *N* é o tamanho da amostra do conjunto de dados e n_p é o número de parâmetros. O segundo termo tem a função de penalizar redes com muitas arestas, o que leva o método BIC a contribuir para geração de grafos mais simples. Para um *N* grande, os modelos com pontuação mais alta geralmente tem parâmetros próximos dos valores de máxima verossimilhança.

Quando existe a necessidade de tolerância maior para redes mais complexas, como por exemplo na fase de análise exploratória, o método *Akaike Information Criterion* (AIC) fornece uma função alternativa de pontuação. Sua formulação é dada por (Eq. 9):

$$AIC = \log\left(\rho\left(D\middle|\hat{\theta}, G\right)\right) - np \qquad (9)$$

O método AIC penaliza menos severamente a inclusão de arestas adicionais e parâmetros associados. É importante observar que a máxima verossimilhança não pode ser usada como função de pontuação, a não inclusão de um termo de penalidade sempre conduziria à seleção de uma rede totalmente conectada.

Su (2013) cita a pontuação K2 de (COOPER; HERSKOVITS, 1992) como um método intermediário entre AIC e BIC que permite extrair uma prévia sobre a estrutura e parâmetros uniformes, cuja função pode ser formulada como (Eq. 10):

$$\log(K2, X_i)) = \sum_{j=1}^{q_i}\left(\ln\left(\frac{(r_i-1)!}{N_{ij}+r_i-1)!}\right) + \sum_{k=1}^{r_i}\ln(N_{ijk}!)\right) \qquad (10)$$

Onde N_{ijk} representa o número de casos na base de dados na qual a variável *Xi* obtém seu k-ésimo valor (k=1, 2,...,r_i), e seu conjunto de pais foi instanciado como sua j-ésima

combinação de valores (j=1,2,..., q_i), e $N_{ijk} = \sum_{k=1}^{r_i} N_{ijk}$. O logaritmo da pontuação total K2 é então a soma das contribuições individuais.

Alguns exemplos desse tipo de algoritmo são (SCUTARI, 2014):

1. *Greddy Search*, *Hill-Climbing* com partidas aleatórias ou *Tabu Search* (BOUCKAERT, 1995), são algoritmos que exploram o espaço de busca iniciando a partir de uma estrutura (normalmente sem arcos) e vai adicionando, removendo e revertendo um arco por vez até que a pontuação não possa mais ser melhorada;
2. Algoritmos Genéticos, que simulam uma evolução natural por meio de seleção interativa de modelos de teste de encaixe (*fittest*) e hibridização de outras características (LARRAÑAGA et al., 1997). Nesse caso o espaço de busca é explorado por meio de operadores estocásticos *crossover* que combinam a estrutura de duas redes e efetuam alterações randômicas;
3. *Simulating Annealing* (BOUCKAERT, 1995), executa uma busca local estocástica, aceitando alterações que aumentam a pontuação da rede e ao mesmo tempo permitem alterações que a diminuem com a probabilidade inversamente proporcional a diminuição da pontuação.

Abordagem híbridas

Uma mistura dos dois métodos, conhecidos como "algoritmos híbridos" foram desenvolvidos com o objetivo de maximizar as suas vantagens e amenizar seus pontos fracos (SCUTARI, 2014). Normalmente, eles iniciam suas atividades por um algoritmo baseado em restrição para encontrar o esqueleto da rede e, em seguida, um método baseado em pontuação é utilizado para identificar o melhor conjunto de arestas (SU, 2013). Entre os algoritmos que utilizam esse método estão (SCUTARI, 2015): o *Sparse Candidate* (SC) (FRIEDMAN et al. 1999) e o *Max-Min Hill-Climbing* (MMHC) (TSAMARDINOS, 2006).

Estes dois algoritmos se baseiam em duas etapas conhecidas como restringe e maximiza. Na primeira etapa o conjunto candidato para pais de cada nó *Xi* é reduzido a partir do conjunto *V* para um conjunto reduzido C_iV de nós, cujo comportamento demonstrou estar relacionado de alguma forma com X_i, o que resulta em um espaço menor e mais regular. A segunda etapa visa a rede que maximiza uma função de pontuação, sujeito às restrições impostas pelo conjuntos C_i. No algoritmo SC esses dois passos são executados

interativamente até que a pontuação da rede não possa ser melhorada. No algoritmo MMHC a função restringe e maximiza são executadas somente uma vez, a heurística *Max-Min Parentsand Children* (MMPC) é utilizada para aprender o conjunto candidato C_i e o algoritmo *Hill-Climbing Greedy Search* para encontrar a rede ótima.

Uma combinação do algoritmo PC e K2 gerou o algoritmo CB Singhy e Valtorta(1994). Este algoritmo utiliza-se de testes de independência condicional gerado pelo PC para construir a ordenação causal das variáveis e a partir dessa ordenação gera a estrutura da rede pelo K2 sem métodos de independência condicional.

O algoritmo BENEDICT (*BElief NEtworks DIscovery using Cut-set Techniques*) (ACID; DE CAMPOS, 2004), que exige a ordenação causal de variáveis, utiliza o conceito de d-separação para calcular a independência condicional entre as variáveis e a entropia cruzada para medir o grau de discrepância entre as estruturas de RBs geradas. A entropia cruzada mede o tanto de dependência existente entre *X* e *Y* dado *Z*. A medida geral de discrepâncias é calculada pela soma dos graus de dependência dos pares formados por nós não adjacentes com bases nos conjuntos de d-separação mínimos gerados. A melhor rede é a que tiver a menor medida de discrepância. A complexidade deste algoritmo é $O(n^6)$.

A próxima seção delineia o funcionamento do método de AT abordando sua definição, terminologias, apresenta regras, conceitua efeito indireto e direto e descreve como se avalia o modelo e intrepreta o modelo.

2.5 ANÁLISE DE TRAJETÓRIAS

Um dos principais objetivos das técnicas multivariadas é ampliar a capacidade explanatória dos pesquisadores e a eficiência estatística. Hair (2005) destacou que a maioria dos métodos estatísticos como: regressão múltipla, análise fatorial, análise multivariada de variância, análise discriminante e outras técnicas fornecem amplo suporte aos pesquisadores científicos. Por outro lado, todas essas técnicas compartilham uma mesma limitação, ou seja, cada uma delas só avalia uma relação por vez, mesmo aquelas técnicas que permitem múltiplas variáveis dependentes. Passados 13 anos dessa afirmação, mesmo com a evolução desses métodos, quando a modelagem exige o padrão VD e VI a maioria dos métodos ainda se limita a avaliar a relação com apenas uma VD. Com exceção dos modelos de séries temporais, todos os outros métodos são mais simples do que a análise de trajetórias (AT) também conhecido como *Path Analysis* (PA). AT é um método que foi desenvolvido por Sewal Wright (1934) como ferramenta para estudar as relações e efeitos diretos e indiretos de

variáveis classificadas (KLINE, 2015). Em resumo, AT permite investigar se *X* influencia *Y* e *Y* influencia *Z*, enquanto que em modelos tradicionais como o de regressão, mesmo a multivariada por exemplo, observa-se se apenas o efeito de *X* em *Y*. Apesar de AT ser o membro mais antigo da modelagem de equações estruturais (MEE), apenas 25% de 500 estudos revisados por MacCallum e Austin (2000) referiam-se a essa técnica.

Neste contexto, nas próximas seções são apresentados alguns dos aspectos fundamentais que caracterizam a AT como um método que fornece uma interpretação quantitativa sobre a relações causais entre as variáveis do modelo.

A seção seguinte apresenta o conceito de AT, suas terminologis e convenções, conceitos considerados importantes para que se inicie o trabalho com este método.

2.5.1 Definição de análise de trajetórias, terminologias e convenções de design

A AT é uma extensão da regressão múltipla que permite examinar relações mais complicadas entre as variáveis, indo além de simplesmente ter VIs prevendo uma VD, e comparar diferentes modelos uns contra os outros para ver qual deles se ajusta melhor sobre um conjunto de dados (STREINER, 2005). Por ser um método, AT não tem o propósito de descobrir causalidades, mas fornecer suporte para pesquisadores na formulação de seus modelos causais (PEDHAZUR, 1997). Nas palavras de Wright:

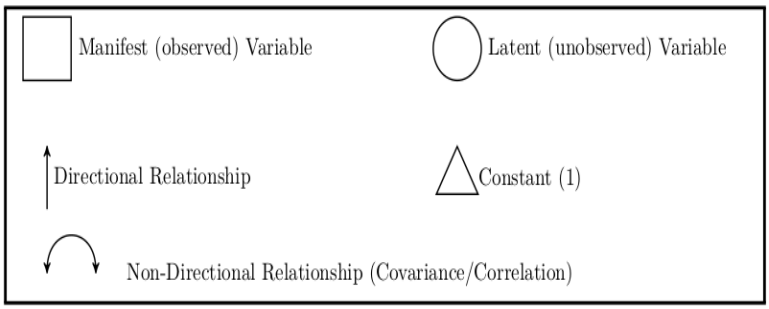
> O método dos coeficientes de trilha não se destina a realizar a tarefa impossível de deduzir relações causais a partir dos valores de coeficientes de correlações. Pretende-se combinar a informação quantitativa dada pela correlação com as informações qualitativas que se possam encontrar sobre as relações causais para dar uma interpretação quantitativa (WRIGHT, 1934, p.193).

Um modelo de AT é uma representação pictórica (diagrama) da teoria que está por trás do relacionamento entre as variáveis. Uma simbologia especial é utilizada para se criar o modelo de AT (BEAUJEAN, 2014) e pode ser observada na Figura 14. Nesta Figura os retângulos representam variáveis observadas (ou manifestas), os círculos ou elipses variáveis não observadas ou latentes (conceito não abordado neste estudo), as setas retas com uma ponta representam a direção do relacionamento (causa para efeito) e influência direta, as setas curvas com duas pontas representam um relacionamento de covariância ou correlação não direcional e o triângulo representa uma constante.

A nomeação de variáveis em AT, considerando a sua contrapartida mais sofisticada a modelagem de equações estruturais (MEE), é diferente da estatística tradicional, justamente para evitar confusão. Ao invés de usar os termos VI e VD, em AT usa-se os termos variáveis

variáveis exógenas (VEX) para aquelas que têm setas retas que emergem delas e nenhuma apontando para elas, exceto quando se usam termos de erro e variáveis endógenas (VEN) que devem ter pelo menos uma seta direta apontando para elas. Esses termos se justificam pelo fato de que as causas ou fatores que influenciam as VEXs são determinados fora do modelo, enquanto que fatores que influenciam VENs estão presentes no próprio modelo.

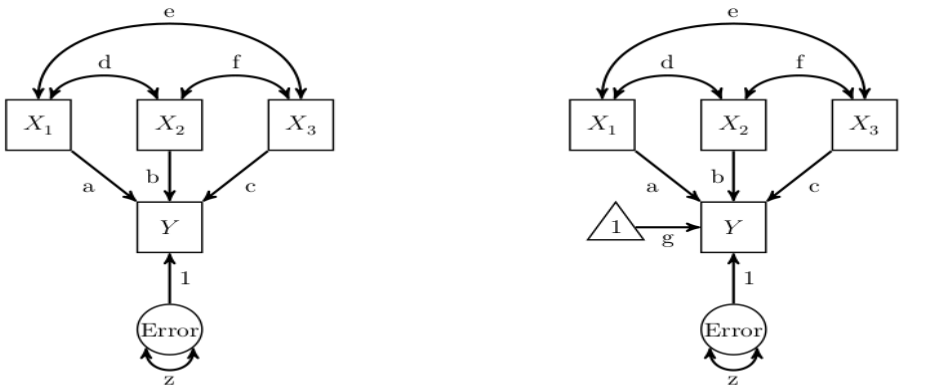
Figura 14 - Símbolos utilizados para construir diagramas de AT.



Fonte: BEAUJEAN, 2014.

A Figura 15 mostra um exemplo de modelo de AT para uma regressão múltipla, onde X_1, X_2 e X_3 são VEXs e Y é VEN. Neste caso, as variáveis X_1, X_2 e X_3 são consideradas ter efeito direto sobre Y e covariar umas com as outras. O triângulo representa o termo de interceptação, ou seja, o valor de Y quando X_1, X_2 e X_3 são iguais a zero e são utilizados apenas em modelos mais complexos. As letras minúsculas sobre cada uma das setas representam o coeficiente do caminho. Este coeficiente pode ser positivo indicando que um aumento na variável causal resultará no aumento do efeito sobre a variável dependente se todas as outras variáveis causais permanecerem constantes. Caso, este coeficiente seja negativo, um aumento na variável causal provocará uma diminuição do efeito sobre a variável dependente (BEAUJEAN, 2014).

Figura 15 - Modelo de AT para uma regressão múltipla com 3 variáveis preditoras (VEXs)



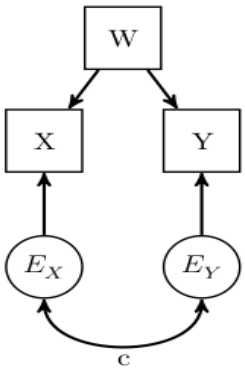
(a) Standardized model: $Y = aX_1 + bX_2 + cX_3 + \text{Error}$.

(b) Unstandardized model: $Y = aX_1 + bX_2 + cX_3 + g + \text{Error}$.

Fonte: BEAUJEAN, 2014

As VENs sempre tem um termo de erro, também conhecido como termo residual ou perturbação, representada pelo círculo associado a ela. Esse termo é similar ao termo de erro inserido no final das equações de regressão. De forma similar à regressão estes capturam duas ocorrências: a) imprecisão na medida de VENs, pois todas as ferramentas de medição sofrem algum grau de erro. b) outros fatores que afetam as VENs e que não foram medidos, seja por falta de tempo, desconhecimento de sua importância, ou outro motivo qualquer. Esse termo representa a discrepância entre valores observados e valores preditos pelo modelo. A variabilidade dessas discrepâncias representa a variação do erro, ou seja, o montante de variância em uma VEN que não está sendo explicada pelas outras variáveis do modelo. Termos de erro são considerados exógenos por representar uma causa direta dentro do modelo. Esses termos ainda podem covariar com outras variáveis. Por exemplo, na Figura 16 as variáveis *X* e *Y* não podem covariar por serem VENs, no entanto seus termos de erro covariam entre si (BEAUJEAN, 2014). Por questões de simplicidade nem sempre essas setas são desenhados, mas devem ser implicitamente considerados.

Figura 16 - Modelo AT de uma correlação parcial



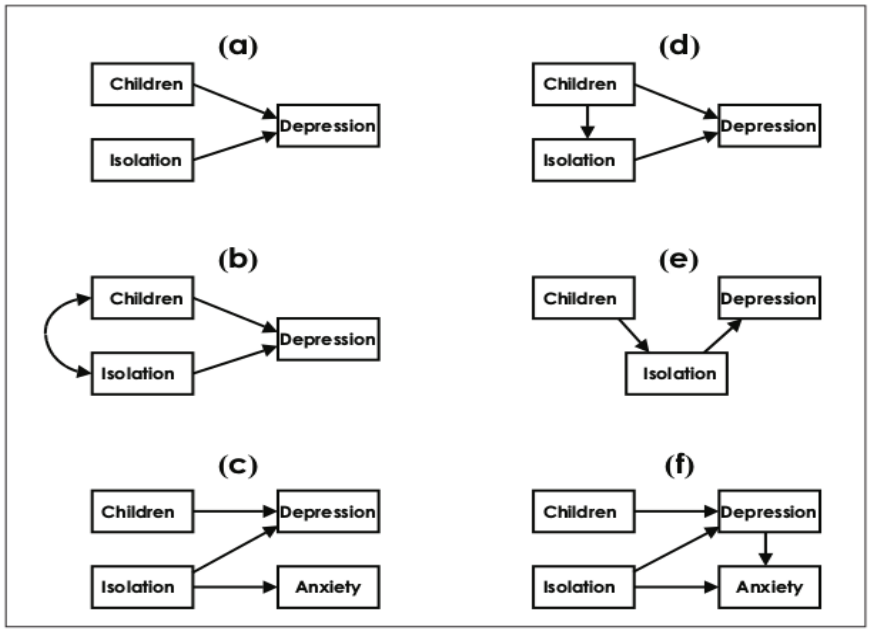
Fonte: BEAUJEAN, 2014.

À medida que se avança em AT outros dois termos importantes que irão aparecer são variância e covariância (STREINER, 2005). Variância é simplesmente a quantidade de variação uma variável para a outra, de maneira formal é o quadrado do desvio padrão (DP). Covariância é considerada primo de primeiro grau da correlação. Correlação informa o comportamento de uma variável em relação a outra, ou seja, se a variável *A* subir então a variável *B* também sobe (correlação positiva) ou baixa (correlação negativa) ou as mudanças são independentes (correlação zero). Durante o cálculo de correlação ambas as variáveis são transformadas em *escores* padrão com média 0 e DP 1, o mesmo se aplica para a covariância, exceto que as variáveis não são transformadas. Em AT o objetivo é determinar se existe

algum padrão significativo entre as variáveis, ou seja, como elas se influenciam e tudo que se tem para resolver esse problema são as covariâncias e correlações. Geralmente, quando se mede várias variáveis em um conjunto de dados, se gera uma matriz que apresenta a variação de cada variável ao longo da diagonal principal e as covariâncias em todas as outras células (STREINER, 2005).

Outros possíveis modelos são representados pela Figura 17. Na Figura17*a* postula-se que as VEXs "filhos" e "isolamento" influenciam a variável "depressão" de forma independente e essas duas variáveis não se correlacionam. Esse é conhecido como modelo independente por refletir a falta de correlação entre as VEXs. Na Figura 17*b*, representa-se um modelo onde as VEXs estão correlacionadas, portanto desenha-se uma seta curva entre elas com duas pontas. Essa seta curva indica correlação ou covariância entre as variáveis. Esse modelo é chamado *run-of-the-mill multiple regression* (sem tradução). Em AT sempre se assume que as VEXs estão correlacionadas e por padrão não se desenha esta seta curva a menos que seja necessário. Na Figura 17*c* amplia-se o modelo e olhando para as duas VENs este indica que "filhos" afeta apenas "depressão" enquanto que "isolamento" influencia tanto depressão quanto "ansiedade". Os modelos apresentados no lado direito da Figura 17 (*d, e, f*) são chamados de mediados ou indiretos porque nele as VEXs atuam sobre uma VEN de forma indireta, ou seja, por meio de sua influência sobre uma outra VEN. Por exemplo, na Figura 17*d* "filhos" influencia diretamente "depressão" e ao mesmo tempo influencia "isolamento" que então influencia depressão. Em outras palavras isolar-se causa depressão, no entanto, a presença de crianças exacerba este efeito tanto direta, como indiretamente. Na Figura 17*e* pode-se interpretar que "isolamento" é causado unicamente por "crianças" e este causa depressão. Finalmente, na Figura 17*f* observa-se que a variável endógena final "depressão" também pode afetar "ansiedade". Isso não limita as possibilidades porque os caminhos podem ser muito mais longos e envolver diversas etapas intermediárias (STREINER, 2005).

Figura 17 - Alguns possíveis modelos de AC



Fonte: STREINER, 2005

A próxima seção descreve o modo como se devem estimar os valores dos coeficientes de caminhos do modelo, processo essencial para se calcular o valor das trajetórias de um modelo de AT.

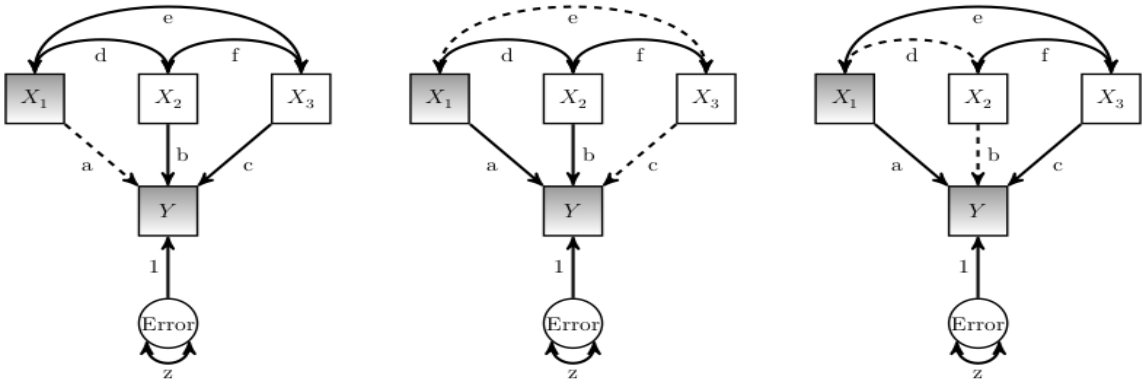
2.5.2 Regras de rastreamento

O geneticista Sewall Wright criou o termo conhecido como “regras de rastreamento” com o objetivo de estimar a covariância entre duas variáveis, ou seja, como estimar os valores para os coeficientes dos caminhos do modelo. Esse processo se dá por meio da varredura dos caminhos dentro do modelo, ou seja, fazendo-se uma análise de trajetórias. Durante esse processo soma-se os caminhos de conexão apropriados. As regras básicas para modelos básicos são (BEAUJEAN, 2014):

1. traçar todos os caminhos entre duas variáveis multiplicando todos os coeficientes ao longo de um determinado caminho;
2. ao avançar ao longo das setas não é possível voltar para trás;
3. laços não são permitidos, ou seja, não se pode passar pela mesma variável mais de uma vez por um caminho já descoberto;
4. no máximo, pode haver uma seta de duas pontas incluída em um caminho;
5. após traçar todos os caminhos para um determinado relacionamento, somar todos os seus valores.

O uso dessas regras são demonstrados pela estimativa do relacionamento entre X_1 e Y na Figura18, representado por "σ_{1Y}". O primeiro passo é encontrar todos os caminhos que permitem ir de X_1 para Y. Neste caso, estes são representados pelos coeficientes de caminho "*a*", "*ec*" e "*db*". Não é possível por exemplo ir pelo caminho "*efda*" por violar ao pressuposto que proíbe laços. Dessa forma a solução para o problema é: *a* + *ec* + *db*. Cálculos similares podem ser feitos para σ_{2Y} e σ_{3Y}.

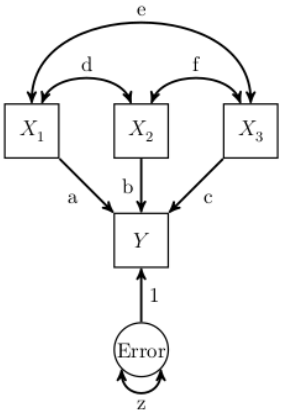
Figura 18 - Modelo de AT demonstrando os possíveis caminhos de X_1 para Y



Fonte: BEAUJEAN, 2014.

Como demonstração, os parâmetros da Figura 19 serão estimados usando-se a Tabela 1 que contém os valores padronizados das covariáveis. Como as variáveis *d*, *e* e *f* são variáveis manifestas (não podem ser medidas, pois seus valores são inferidos de outras medições), sabemos seus valores: *d* = 0.20, *e* = 0.24 e *f* = 0.30.

Figura 19 – Modelo de trajetórias de uma regressão múltipla com três variáveis preditoras (exógenas)



(a) Standardized model:
$Y = aX_1 + bX_2 + cX_3 + \text{Error}.$

Fonte: BEAUJEAN, 2014.

Tabela 1 – Valores padronizados das covariáveis de Y

	X_1	X_2	X_3	Y
X_1	1,00			
X_2	0,20	1,00		
X_3	0,24	0,30	1,00	
Y	0,70	0,80	0,30	1,00

Fonte: o autor, 2018.

Para resolver o problema, aconselha-se escrever o conjunto que compõe cada caminho para as correlações entre a VEN Y e as VEXs X_1, X_2 e X_3 e então substituir os valores:

$$r_{1Y} = a + ec + db \rightarrow 0.70 = a + 0.24c + 0.20b$$
$$r_{2Y} = b + da + fc \rightarrow 0.80 = b + 0.20a + 0.30c$$
$$r_{3Y} = c + fb + ea \rightarrow 0.30 = c + 0.30b + 0.24a$$

Por meio de um pouco de álgebra a resolução resulta em: $a = 0.57$, $b = 0.70$, $c = -0.05$.

Para confirmar os resultados, os valores estimados para a,b e c foram substituídos em uma das equações originais, no caso r_{1Y}para averiguar se o resultado retornado está correto.

$$r_{1Y} = 0.70 = \quad a + 0.24c + 0.20b$$
$$0.70 = 0.57 + 0.24 * (-0.05) + 0.20 * (0.70)$$
$$0.70 = 0.57 - 0.01 + 0.14$$
$$0.70 = 0.70$$

A resolução de z foi construída seguindo-se o mesmo raciocínio, observando que $r_{YY} = \sigma^2_{YY} = 1$ umas vez que as variáveis são padronizadas.

$$r_{YY} = a^2 + b^2 + c^2 + 2(cea) + 2(cfb) + 2(bda) + z$$
$$1,00 = 0,57^2 + 0,70^2 + -0,05^2 + 2(-0,05)(0,24)(0,57) +$$
$$2(-0.05)(0.30)(0.70) + 2(0.70)(0.20)(0.57) + z$$
$$1.00 = 0.335 + 0.490 + 0.002 + -0.014 + -0.021 + 0.160 + z$$
$$1.00 - 0.952 = z$$
$$0.048 = z$$

Para estimar a quantidade de variância de *Y* que as variáveis preditoras explicam (R^2), usa-se a mesma fórmula para estimar r_{YY}, porém omite-se o *z*.

$R^2 = a^2 + b^2 + c^2 + 2(cea) + 2(cfb) + 2(bda) = 0.942$

$R^2 = 0{,}57^2 + 0{,}70^2 + -0{,}05^2 + 2(-0{,}05)(0{,}24)(0{,}57) +$

$2(-0.05)(0.30)(0.70) + 2(0.70)(0.20)(0.57)$

$R^2 = 0.335 + 0.490 + 0.002 + -0.014 + -0.021 + 0.160 = 0.952$

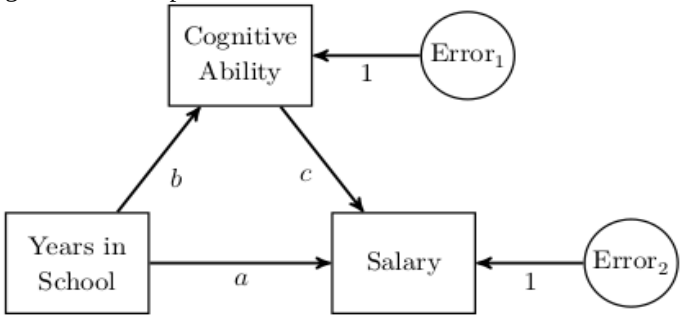
Observando-se a Figura 19, os valores das variáveis *d* e *f* são coeficientes de correlação e os valores *a, b, c* são coeficientes de regressão parcial (significa que é a relação entre uma variável exógena e endógena, controlando para todas as outras variáveis exógenas que vão para aquela variável endógena) padronizados (variáveis transformadas em escores *z* em unidades de desvio padrão), também conhecidos como coeficientes de caminho (*path coefficients*).

A seguir se apresenta o conceito relativo a efeito indireto, um conceito possível de modelar com e metodologia de AT.

2.5.3 Efeito indireto

Efeitos indiretos ocorrem quando existe a influência de uma variável sobre outra por meio de uma terceira ou quarta variável (BEAUJEAN, 2014). A Figura 20 apresenta um exemplo de AT com efeito indireto, neste caso, se postula que no modelo existe uma influência direta da variável "*anos na escola*" sobre a variável "*salário*", representada pelo caminho "*a*". Adicionalmente, também se postula que há uma influência indireta de "*anos na escola*" sobre a variável "*salário*", mas passando pela variável "*capacidade cognitiva*", representada pelos caminhos "*b*" e "*c*".

Figura 20 - Exemplo de um modelo de AT com efeito indireto.



Fonte: BEAUJEAN, 2014

Dessa maneira, seguindo-se as "regras de rastreamento" (seção 2.5.2) o caminho de "*anos na escola*" para "*salário*" passando por "*capacidade cognitiva*" é calculado por *b*c = bc* (efeito indireto) e por *a* (efeito direto).

Considerando-se um outro exemplo para visualizar o efeito indireto e a flexibilidade de AT, suponha-se que se deseja saber se a pontuação de uma pessoa com fotonumerofobia (PNP) (medo de que números apareçam) pode ser predita por 3 fatores: nível global de ansiedade (ANX), grau de de conhecimento de matemática no ensino médio (HSM) e a discrepância entre o que a pessoa estimou de impostos no ano passado a pagar para o imposto de renda (IR) e o que realmente era (TAX). Neste caso, estamos postulando que a fotonumerofobia se correlacione de forma positiva com a ansiedade de uma pessoa (aumentando-a), de forma negativa em suas aulas de matemática (atrapalhando a aprendizagem) e de forma positiva a confusões com cálculo de impostos ao declarar seu IR (induzindo a erros). A formulação deste problema em forma de regressão seria (STREINER, 2005) (Eq.11):

$$PNP = b_0 + b_1ANX + b_2HSM + b_3TAX + Error \qquad (11)$$

Onde, b_0 é o intercepto e os outros *bs* são os declives da reta de regressão para cada VI. O cálculo das correlações, as médias e desvios padrões das variáveis e suas respectivas correlações são apresentadas na Tabela 2. Neste exemplo, o cálculo da correlação, representado por *R* é de 0,638 e o R^2, que reflete a proporção de variância da VD, é de 0,407.

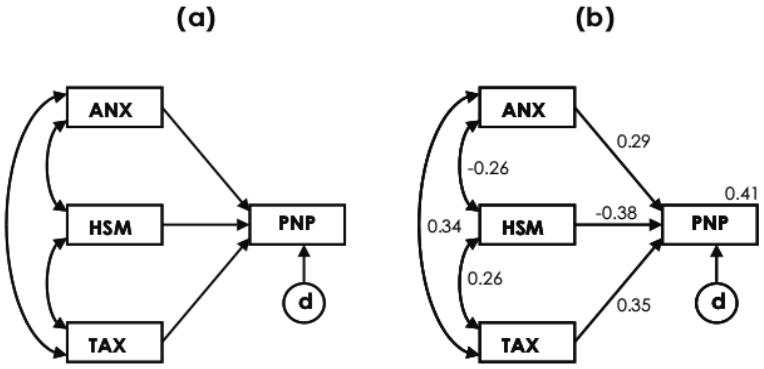
Tabela 2 - Média, DP e correlações entre as variáveis e pesos de regressão padronizados (b) e não padronizados (β) para a relação entre fotonumerofobia (PNP), ansiedade (ANX) grau de de conhecimento de matemática no ensino médio (HSM) e discrepância na taxa de imposto de renda (TAX)

	PNP	**ANX**	**HSM**	**TAX**	**Média**	**DP**	**b**	**β**
PNP	1.000	0,509	-0,366	0,346	26,79	7,33		
ANX		1.000	-0,264	0,338	20,33	5,17	0,414	0,292
HSM			1.000	0,260	74,69	5,37	-0,517	-0,379
TAX				1.000	1983,23	535,49	0,005	0,346

Fonte: STREINER, 2005.

O resultado desta regressão pode ser observado no grafo de AT apresentado pela Figura 21, onde se observa o valor das correlações entre os preditores (ao lado das setas curvas), os pesos de *b* ao lado das setas retas e o R^2 na caixa representando PNP. Do ponto de vista teórico, o grafo de AT corresponde ao modelo teórico.

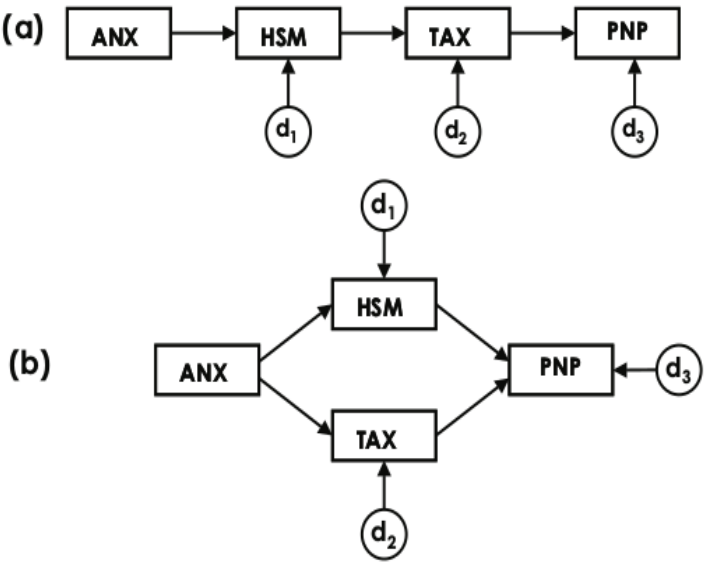
Figura 21 - Modelo hipotetizado (a) e resultados (b) do modelo de regressão



Fonte: STREINER, 2005.

Executar cálculos de regressão múltipla é trivial para muitos programas de computador. Para muitos é desnecessário usar *software* especial de AT para resolver este tipo de problema. No entanto, AT pode oferecer muito mais, pois essa técnica permite postularmos outras hipóteses sobre as relações entre as variáveis e avaliar se estas geram uma contabilização de variância melhor ou pior em PNP. Por exemplo, a Figura 22*a* hipotetiza-se que ao invés das variáveis atuarem de forma separada em PNP, ansiedade leva a piora do aprendizado em matemática, o que gera erros no cálculo e preenchimento dos formulários do IR que consequentemente leva a fotonumerofobia. Considerando-se a Figura 22*b* hipotetiza-se que ansiedade resulta ao mesmo tempo em piora do aprendizado em matemática e erros no cálculo e preenchimento do IR e que ambos implicam em fotonumerofobia. Essa última hipótese caracteriza o efeito indireto, representado por duas trajetórias de ansiedade (*ANX-HSM* e *ANX-TAX*), que podem contribuir como efeito sobre fotonumerofobia.

Figura 22 - Outros possíveis modelos



Fonte: STREINER, 2005.

A avaliação da qualidade do ajustamento do modelo tem o objetivo de avaliar quão bem o modelo teórico é capaz de reproduzir a estrutura correlacional das variáveis observadas na amostra sob estudo. A seção a seguir descreve como executar este processo.

2.5.4 Avaliação da qualidade do modelo

Essa é uma das áreas menos consensuais em MEE (MAROCO, 2010; KLINE, 2015), pois diversos estudos (BARRET, 2007; BENTLER, 1990; BOLLEN; LONG, 1992; BROWNE, et al., 1993; MCINTOSH, 2007; MULAIK, 2007) apresentam simulações e observações empíricas que levam a justificar diferentes estratégias e recomendações para análise da qualidade do ajustamento.

O processo de avaliação da qualidade do modelo geralmente é feito com 3 tipos de testes: (*i*) testes de ajustamento, (*ii*) avaliação de índices empíricos cuja base se origina nas funções de verossimilhança ou na matriz de resíduos obtidos durante o ajuste do modelo e (iii) a análise de resíduos e da significância dos parâmetros. Na próxima seção serão apresentadas as principais estatísticas e índices reportados na maioria dos estudos e aplicações (MAROCO, 2010):

Teste de ajustamento

O teste qui-quadrado ou X^2 é responsável pela avaliação da função de discrepância do modelo minimizado após o seu ajustamento. Durante esta fase avalia-se as hipóteses estatísticas do modelo que podem ser:

H_0: $\Sigma = \Sigma(\hat{\theta})$ A matriz de covariância populacional é igual a matriz de covariância estimada pelo modelo.

H_1: $\Sigma \neq \Sigma(\hat{\theta})$ A matriz de covariância populacional não é igual a matriz de covariância estimada pelo modelo.

A estatística do teste é calculada por (BOLLEN, 1986; JORESKOG;SORBOM, 1996) (Eq. 12):

$$X^2 = (n - 1) f_{min} \sim x^2(g.l.) \qquad (12)$$

Onde f_{min} é o valor mínimo de uma das funções de discrepâncias para o método ML, GLS ou WLS. A maioria dos *software* de MEE apresenta o p-valor do teste calculado como $p - valor = 1 - \Phi(X^2, \text{gl})$ onde Φé a função de distribuição do X^2. Para um determinado nível de significância (α) rejeita-se H_0 se p-valor <=α. Quanto maior for o valor do X^2pior será o ajustamento. O teste do qui-quadrado exige alguns cuidados por ser altamente sensível ao tamanho da amostra, pois em amostras grandes ocorre a possibilidade de rejeitar a hipótese de que o modelo se ajusta bem aos dados quando o ajuste é bom (erro tipo I) e em amostras pequenas ocorre o contrário, ou seja, o teste não rejeita a hipótese de que o modelo se ajusta bem aos dados quando o ajuste é ruim (erro tipo II) (MAROCO, 2010).

O teste do qui-quadrado é sensível à violação dos pressupostos sobre a distribuição normal das variáveis. Para corrigir esse problema criou-se a "correção de Satorra-Bentler", que introduz na estatística do teste uma função de curtose multivariada amostral, o tipo de modelo e o método de estimação, reduzindo de forma considerável a probabilidade de erro tipo I quando o pressuposto de normalidade multivariada não é válido Satorra e Bentler, (2001). A sua estatística é calculada por (Eq. 13):

$$X^2_{SB} = \frac{X^2}{c} \sim x^2(gl) \qquad (13)$$

Onde X^2 é a estatística e *c* é um fator de correção.

Índices de qualidade do ajustamento

Em virtude de algumas limitações e problemas do teste do qui-quadrado, foram desenvolvidos alguns índices de avaliação de qualidade do ajuste do modelo, conhecidos como "*goodness-of-fit*". A ideia principal desses índices é quantificar a qualidade do ajustamento do modelo em face a modelos de referência que avaliam o melhor ajuste possível, dito 'modelo saturado' no qual todas as trajetórias correlacionais entre as variáveis possuem qui-quadrado = 0, ou com o modelo de pior ajuste, dito 'modelo de independência total' onde se considera que nenhuma variável está correlacionada com as variáveis restantes atingindo o valor máximo do qui-quadrado. Esses índices podem ser utilizados com alternativas ao qui-quadrado e são classificados em 5 famílias, descritas a seguir:

Índices Absolutos: esses índices avaliam a qualidade do modelo por si, sem compará-los com outros modelos. Os índices mais comuns são:

a. $X^2/g.l.$: Se H_0 do teste do qui-quadrado de ajustamento for verdadeira, se espera que o valor referente aos graus de liberdade seja igual ao valor esperado da estatística do teste. Portanto, para um ajustamento perfeito o resultado será igual a 1. No entanto, considera-se bom se o resultado for inferior a 2, aceitável se inferior a 5 e inaceitável para valores superiores (ARBUCKLE, 2008).

b. *Root Mean Square Residual* (RMR): É a raiz quadrada da matriz de erros dividida pelos graus de liberdade assumindo-se que o modelo ajustado seja o correto Joreskog e Sorbom (1989). RMR próximos de zero indicam um ajustamento melhor, sendo RMR = 0 perfeito (Eq. 14):

$$RMR = \sqrt{\frac{\sum_{q}^{p} \sum_{j=1}^{i} (S_{ij} - \sigma(\theta)^2}{(p+q)+(p+q=1)/2}} \qquad (14)$$

B. Índices Relativos: Executam a avaliação do modelo comparando-o com (i) modelo de independência total e/ou (ii) modelo saturado:

a. *Normal Fit Index (NFI*): Este índice foi proposto por Bentler e Bonett (1980) e avalia o percentual de aumento da qualidade do ajuste do modelo ajustado (X^2) em relação ao modelo de independência total ou modelo basal (X_b^2) (Eq.15):

$$NFI = 1 - X^2 / X_b^2 \qquad (15)$$

NFI inferior a 0.8 indica que o modelo ajustado está a 80% do percurso entre o pior modelo e o melhor modelo possível e indicando um mau ajuste, valores entre 0.8 e 0.9 indicam um ajuste sofrível e acima de 0.9 um bom ajuste. NFI igual a 1 indica um ajuste perfeito (ARBUCKLE, 2008). Quanto maior o número de variáveis do modelo ou a dimensão da amostra mais elevado se torna o valor de NFI.

b. *Comparative Fit Index* (CFI): Proposto por Bentler (1990) foi elaborado para corrigir problemas quando se usa o NFI com amostras pequenas. O CFI compara o ajustamento do modelo em estudo (X^2) com graus de liberdade gl , com o ajustamento do modelo basal (X_b^2) com graus de liberdade gl_i (Eq. 16)

$$CFI = 1 - \frac{max\,(X^2) - gl,0)}{max\,(X_b^2 - gl_b,0)} \qquad (16)$$

CFI com valores inferiores a 0.9 sinalizam um mal ajuste, entre 0.9 e 0.95 um ajuste bom e acima de 0.95 indicam um ajuste muito bom, e igual a 1 um ajuste perfeito. Este índice independe do tamanho da amostra, mas conforme se aumenta o número de variáveis se diminui o CFI.

c. *Relative Fit Index (RFI)*: Este índice avalia o ajuste do modelo comparando o X^2 normalizado, pelos graus de liberdade, com o modelo basal (Eq.17):

$$RFI = 1 - \frac{X^2 / gl}{X_b^2 gl_b} \qquad (17)$$

Valores de RFI próximos de 1 indicam um bom ajustamento, valores inferiores a 0.9 indicam mal ajuste (BOLLEN, 1996). Este índice apresenta o mesmo problema do NFI.

d. *Tucker-Lewis Index (TLI)*: Conhecido também por *Bentler-Bonnet nor-normed fit index* (NNFI). Este índice é definido por Bentler e Bonnet (1980) como (Eq.18):

$$TLI = \frac{\frac{X_b^2}{gl_b} - \frac{X^2}{gl}}{\frac{X_b^2}{gl_b} - 1} \qquad (18)$$

Onde X^2 e X_b^2 com seus respectivos graus de liberdade (gl e gl_B) são definidos na mesma forma que no CFI. Valores de TLI próximo de 1 indicam um ajuste muito bom.

C. Índices de Parcimônia: São obtidos por meio da correção dos índices relativos com um fator de penalização associado à complexidade do modelo. Este índice tem o objetivo de compensar a melhoria ‘artificial’ do modelo proposto por meio da inclusão de mais parâmetros livres de modo que se aproxime este modelo do modelo saturado. Os modelos de maior complexidade podem ter melhor ajuste do que os mais simples (parcimoniosos), no entanto, podem não ser generalizáveis para outras amostras (MULAIK et al., 1989). Por esse motivo, esse tipo de índice penalizam os índices relativos por um fator de complexidade ou relação de parcimônia estimada por gl / gl_b.

   a. *Parsimony CFI (PCFI): Este índice penaliza o CFI pela relação de parcimônia* (Eq. 19)*:*

$$PCFI = CFI \times gl \div gl_b \qquad (19)$$

   b. *Parsimony GFI (PGFI): penaliza o GFI pela relação de parcimônia* (Eq. 20)*:*

$$PGFI = GFI \times gl \div gl_b \qquad (20)$$

   c. *Parsimony NFI (PNFI): penaliza o NFI pela relação de parcimônia* (Eq.21)*:*

$$PNFI = NFI \times gl \div gl_b \qquad (21)$$

   Para todos os índices de parcimônia considera-se que valores inferiores a 0.6 indicam um mau ajuste, valores entre 0.6 e 0.8 indicam um ajuste razoável e valores acima de 0.8 indicam um bom ajuste.

D. Índices de discrepância populacional: Estes índices avaliam se o modelo ajustado é aproximadamente correto comparando o ajuste obtido na amostra com o ajuste que seria obtido se o mínimo da função de discrepância fosse obtido a partir de momentos populacionais.

a. Parâmetroda não Centralidade (NCP):O NCP estima quão afastado se encontra o valor esperado da estatística X^2, sob a validade da H_0, do verdadeiro valor de X^2. A estatística deste parâmetro é estimada por (Eq. 22):

$$NCP = max[X^2 - gl, 0] \qquad (22)$$

O NCP reflete o grau de desajuste do modelo proposto à estrutura da variância-covariância observada. Quanto mais próximo de zero, melhor é o ajuste do NCP.

a. F_0: Essa estatística é o mínimo relativo do NCP (Eq.23):

$$F_0 = max[(X^2 - gl) \,/\, (n-1), 0] = \frac{NCP}{n-1} \qquad (23)$$

Quanto mais próximo de zero for F_0 melhor será o ajuste do modelo.

c. Root Mean Square Error of Approximation (RMSEA): O índice F_0 tende a favorecer modelos mais complexos por estes apresentarem maior número de parâmetros e consequentemente melhor ajuste que modelos mais simples. Steiger et al. (1990) propuseram a penalização do F_0 pelo número de graus de liberdade do modelo para compensar esse potencial favorecimento. A sua estatística é dada por (Eq. 24):

$$RMSEA = \sqrt{F_0 \,/\, gl} \qquad (24)$$

O ajuste do modelo é considerado ruim quando o valor do RMSEA é superior a 0.10, entre 0.08 e 0.10 é considerado medíocre, entre 0.05 e 0.08 bom e inferior a 0.05 muito bom (ARBUCKLE, 2008). No entanto a adição de mais variáveis pode inflacionar o valor do RMSEA.

E. Índices baseado na informação: Esses índices se baseiam na estatística do qui-quadrado e penalizam o modelo em virtude da sua complexidade. Esses índices não possuem valores de referência para classificar o ajuste do modelo como bom

ou ruim. Na verdade, estes são utilizados quando é necessário comparar modelos. Neste caso, o melhor modelo será aquele que apresentar os menores valores em um ou mais desses índices descritos a seguir:

a. Akaike Information Criterion (AIC): O critério de informação AIC é fornecido por (ARBUCKLE, 2008) pela estatística (Eq. 25):

$$AIC = X^2 + 2t \qquad (25)$$

Onde t representa o número de parâmetros estimados pelo modelo.

b. Browne-Cudeck Criterion (BCC): Este critério de informação também fornecido por (ARBUCKLE, 2008) é calculado por (Eq. 26):

$$BCC = X^2 + 2t \frac{\frac{(n-1)[(p+q)(p+q+3)]}{n-(p+q)-2}}{(p+q)(p+q+3)} \qquad (26)$$

Onde p e q, representam o ńumero de variáveis dependentes e independentes exógenas do modelo. Se comparado ao AIC o BCC penaliza mais os modelos complexos.

c. Bayes Information Criterion (BIC): O critério de informações de Bayes é calculado como (Eq. 27):

$$BIC = X^2 + tLn(n) \qquad (27)$$

Se comparado ao critérios de informações AIC e BCC o BIC penaliza mais os modelos complexos, portanto este tende a favorecer os modelos mais simples.

d. Expected Cross-Validation Index (ECVI): Este indice reflete o ajuste teórico do modelo em outras amostras semelhantes àquelas em que o modelo foi ajustado, a partir de uma única amostra. Sua estatística é calculada pela expressão (ARBUCKLE, 2008) (Eq. 28):

$$ECVI = \frac{AIC}{n-1} \qquad (28)$$

Caso o método de estimação seja ML deve-se substituir o ECVI pelo MECVI (Eq. 29):

$$MECVI = \frac{1}{n}\text{BCC} \qquad (29)$$

Tanto o ECVI como o MECVI devem ser utilizados para comparar modelos não aninhados. O modelo mais estável é aquele com menor valor.

Os diversos índices apresentados compõem um conjunto dos mais usados na literatura para avaliar o ajuste do modelo em MEE. Não é comum reportar todos os índices, pois alguns são redundantes, portanto surge a dúvida de qual índice utilizar. Neste contexto, o Quadro2 apresenta as estatísticas e índices mais utilizados pela maioria dos autores de MEE.

Quadro 2 - Estatísticas e índices de ajustes mais utilizados na literatura sobre MEE

Estatística	Valores de Referência
X^2 e p-valor	Quanto menor, melhor; p > 0.05
X^2 / gl	> 5 - ajuste ruim 2 a 5 - ajuste sofrível 1 a 2 - ajuste bom ~ 1 - ajuste muito bom
CFI GFI TLI	< 0.8 - ajute ruim 0.8 a 0.9 - ajuste sofrível 0.9 a 0.95 - ajuste bom >- 0.95 - ajuste muito bom
RMSEA (I.C. 90%) e p-valor (H0: rmsea <= 0.05)	> 0.10 - ajuste ruim 0.05 a 0.10 - ajuste bom <= 0.05 - ajuste muito bom p-valor >= 0.05
AIC BCC ECVI MECVI	Adequados para comparar modelos (especialmente modelos não aninhados) Quanto menor, melhor será o ajuste

Fonte: MAROCO, 2010.

Análise de resíduos e significância dos parâmetros

O ajuste do modelo aos dados pode ser ‘global’, ou seja, se calculam medidas do ajuste global médio aos dados. Este ajuste global pode apresentar-se como ‘bom’ porém uma ajuste local ‘ruim’ pode existir, neste caso um ou mais parâmetros do modelo podem apresentar valores não significativos ou possuir a confiabilidade de um ou mais indicadores reduzida. Portanto se faz necessário diagnosticar esses possíveis problemas, o que é possível por meio das seguintes estatísticas:

a. Avaliação dos resíduos padronizados: Resíduos com valor absoluto superior a 2 indicam com 95% de confiança, que as observações são muito distantes das outras observações (*outliers*) e portanto indicam problemas de ajuste local;
b. Avaliação dos erros-padrão assintóticos dos parâmetros do modelo e sua significância: Erros padrão superior a 2x a estimativa do parâmetro indicam problemas com a estimativa desse parâmetro (multicolinearidade, *outliers* ou subamostragem). Parâmetros não significativos sugerem que existe problema na especificação do modelo. Neste caso rejeita-se H_0 se o p-valor do teste for inferior ou igual a α;
c. Avaliação da confiabilidade individual das variáveis: Essa confiabilidade é estimada por meio da fração da variância de determinada variável que é explicada pelo fator latente. É um conceito similar ao R^2, onde os softwares de MEE calculam um R^2 para cada variável endógena. Valores de R^2 inferiores a 0,25, indicam que o fator explica menos de 25% da variância da variável, indicando possíveis problemas de ajuste local com esta variável.

Conforme visto, algumas estratégias de avaliação de qualidade do ajuste do modelo podem ser falhas, ou apontar um bom ajuste global quando há problemas no ajuste local. Portanto a estratégia de avaliação do modelo deve ser composta por várias medidas de ajuste tanto global e local. Neste contexto, se todas apresentarem um resultado positivo, o pesquisador pode concluir que o seu modelo reproduz, de forma conveniente, a estrutura relacional existente entre as variáveis (MAROCO, 2010).

Erros que resultam em modelos com coeficientes de trajetória baixos e não significativos, indicando que o modelo deve ser repensado podem acontecer por falta de

conhecimento do pesquisador. Portanto a seção subseqüente apresenta importantes informações sobre a especificação e identificação do modelo.

2.5.5 A especificação e identificação do modelo

Um modelo pode ser considerado como mal ajustado por diversos motivos. O mais provável é a falta de especificação coerente do modelo, que pode ser gerada por vários motivos, como por exemplo: a inclusão de variáveis que não se relacionam com nenhuma variável endógena, a ausência de variáveis essenciais ao modelo, a criação de trajetórias variáveis que de fato não se relacionam. (STREINER, 2005).

Considerando-se o modelo representado pela Figura 21*b* se observará resultados incomuns, neste caso, todos os índices de ajustes irão apresentar valores acima de 0,90 (recomendado) indicando um ajuste perfeito e o qui-quadrado será tão baixo quanto possível: 0,00. Embora este modelo faça sentido, não existe modelo tão bom, pois o senso comum indica que as variáveis de um modelo sempre são medidas com algum grau de erro. Neste contexto, mesmo um modelo teórico refletindo perfeitamente a realidade este não apresentaria uma correspondência entre este e os dados de forma tão perfeita. Voltando à Figura 21*b* o R^2 é apenas 0,41, o que não faz sentido pois nesse caso o modelo "perfeito" não corresponde à maioria da variância em PNP. Este problema está relacionado à identificação do modelo que se refere à quantidade de parâmetros a se estimar em relação à quantidade de informação que se pode derivar a partir dos dados em termos de variância e covariância entre as variáveis (STREINER, 2005). A identificação do modelo pode se classificar em três categorias: a) Apenas identificado ou "*just identified*" que ocorre quando a quantidade de informação (parâmetros) é exatamente igual ao número de trajetórias que temos para estimar e nesse caso os graus de liberdade (GL) são iguais a zero, b) sub-identificado ou "*under identified"*, neste caso a quantidade de informações é menor ou c) super-identificado, quando a quantidade de informações é maior (BEAUJEAN, 2014; KLINE, 2015).

Para que fique mais claro, considera-se a necessidade de descobrir quais são os valores dos termos da equação 30 (STREINER, 2005) (Eq.30):

$$A + B = 10 \qquad (30)$$

Neste caso há infinitas possibilidades (A=10 e B=0, A=9 e B=1, etc), não sendo possível determinar uma única resposta, ou seja, neste caso não existe informação suficiente

para resolver esta equação e não é possível derivar valores únicos para as duas incógnitas nem calcular os vários parâmetros do modelo, isso identifica um modelo “sub-identificado”. Caso se conheça uma das variáveis, por exemplo que A = 7 a resolução do problema é única, ou seja, B=3, neste caso o modelo é “apenas identificado”. Em uma situação onde existem mais informações do que o necessário proporcionando testar diferentes hipóteses por meio de modelos diferentes uns contra os outros tem-se o modelo considerado “super identificado”. Por fim, é importante destacar que não se pode obter índices de modelos apenas identificados, mas se podecom modelos super-identificados (STREINER, 2005).

Para descobrir quantos parâmetros se pode estimar em um modelo, basta observar o número de variáveis deste e se calcula dessa forma: um modelo com *k* variáveis resulta ([k^2 + *k*] / 2) informações. Como exemplo prático a Figura 21 apresenta 4 variáveis e permite estimar um máximo de ([4^2 + 4] / 2) = 10 parâmetros. Observando-se os 3 possíveis caminhos partindo das VEX (ANX, HSM e TAX) para a VEN (PNP) observa-te também a existência de 3 covariâncias (ou correlações) entre as VEXs. Existe a variâncias dos termos de perturbação (*d*) e finalmente as variações próprias das 3 VEXs, totalizando 10 parâmetros. Os graus de liberdade são compostos pela diferença entre o número de parâmetros que se pode estimar e o número de parâmetros que se deseja estimar (STREINER, 2005).

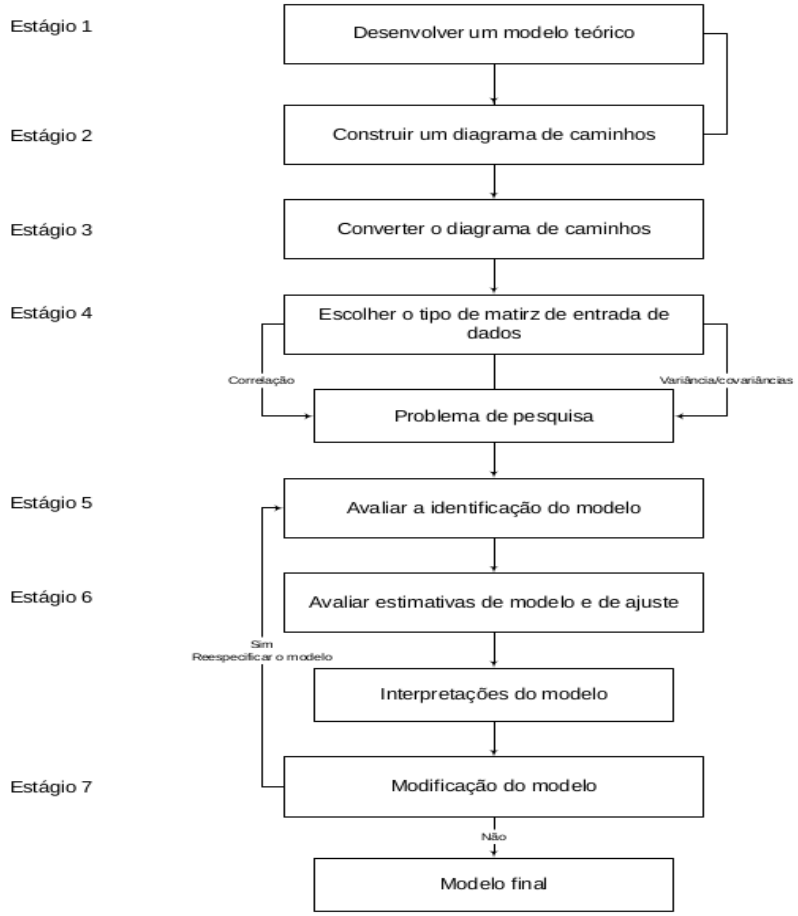
2.5.6 A construção do modelo

Para construção de um modelo de MEE, área da qual modelagem de AT faz parte, Hair et al.(1998) recomenda que se organize o processo em sete etapas. Outros autores como Kline (1998) e Iriondo, Albert e Escudero(2003) corroboram com essa organização de etapas, embora omitam uma delas. Essas etapas são apresentadas pela Figura 23. Neste contexto Streiner (2005) destaca que a construção do modelo deve ser realizada com parcimônia, ou seja, não se deve inserir qualquer variável no modelo e tentar todas as possibilidades de caminhos possíveis para ver o resultado final. Ao contrário, este deve ter uma base sólida preferencialmente apoiada na teoria que a literatura apresenta.

A modelagem deste tipo de modelo se baseia nas relações causais, ou sejam a mudança em uma variável resultará na mudança em outra variável. Ao elaborar este modelo o pesquisador deve assumir que existe causalidade por meio de uma justificativa teórica, obtida por uma pesquisa prévia, que forneça suporte às suas afirmações no modelo. De acordo com Kline (1998) e Hairet al.(1998) existem 4 critérios que devem ser atendidos para que se estabeleça a relação causal entre duas variáveis *X* e *Y*:

1. deve existir precedência cronológica entre as variáveis, ou seja, a variável *X* como causador de *Y* deve precedê-la no decorrer do tempo;
2. a direção da relação causal deve ser especificada de forma correta, ou seja, ao invés de *Y* ser causador de *X* ou ambos exercerem causalidade mútua, *X* deve ser a causa de *Y*;
3. a relação entre *X* e *Y* não desaparece quando variáveis externas como as suas causas comuns se mantêm-se constantes;
4. que exista uma base teórica para a relação.

Figura 23 - Etapas para construção de um modelo de MEE



Fonte: HAIR, 1998.

Esta é a etapa mais difícil e importante da modelagem de AT, pois todas as etapas posteriores assumem que o modelo é correto. Recomenda-se que o pesquisador tenha em mente uma lista de possíveis alterações do modelo justificadas de acordo coma teoria ou resultados empíricos. Essa alteração pode ser necessária caso o modelo não tenha um ajuste

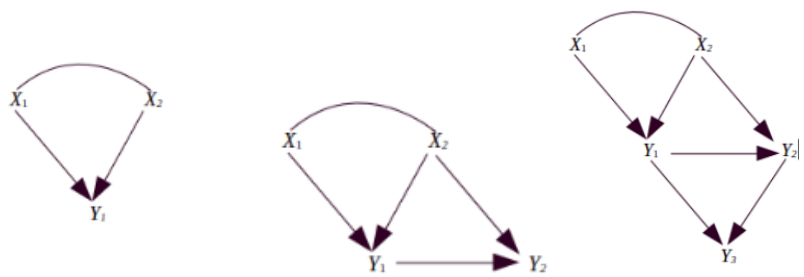
satisfatório e a sua respecificação deve respeitar os mesmos princípios da especificação (KLINE, 2015).

Construir um diagrama de trajetórias

Nesta fase o pesquisador deve representar suas hipóteses na forma de um modelo de AT. Inicialmente se especifica o modelo pelo desenho de um diagrama de AT do modelo por meio de símbolos gráficos padrões. O modelo pode também ser representado por uma série de equações que definem os parâmetros do modelos, os quais correspondem às relações presumidamente existentes entre as variáveis observadas (IRIONDO; ALBERT; ESCUDERO, 2003; KLINE, 2015). O diagrama de AT deve ser construído com base em conhecimento a priori de relações causais relacionadas a experiências anteriores ou base teórica do próprio pesquisador (BEAUJEAN, 2014; KLINE; 2015). Duas suposições são estabelecidas em diagramas de AT, a primeira dita que todas as relações causais devem ser baseadas na teoria e a segunda que as relações causais devem ser assumidas como lineares. No entanto, como em técnicas multivariadas, modelos modificados podem compatibilizar o uso desse tipo de relações (HAIR, 1998).

Após a construção do diagrama de trajetórias, deve se especificar o modelo em termos mais formais. Durante este processo se apresenta uma série de equações que estabelecem: a) as equações estruturais que conectam as variáveis; b) o modelo de mensuração especificando quais variáveis são medidas; e c) matrizes que indicam as correlações teorizadas entre as variáveis. Essas equações resultaram nos parâmetros do modelo, as quais correspondem às relações causais entre as VEXs e VENs do modelo calculadas por meio de dados amostrais processados pelo *software*de AT (KLINE, 1998). Traduzir um modelo de AT em uma série de equações é um procedimento onde cada VEN pode ser prevista por uma ou mais VEXs ou outras VENs (HAIR et al., 1998). A Figura 24 representa 3 modelos diferentes e o Quadro 3 ilustra o processo de tradução para estes modelos de AT.

Figura 24 - Relações causais representadas por meio de diagramas de trajetórias



Fonte: HAIRet al., 1998.

Quadro 3 - Tradução do diagrama de trajetórias para equações estruturais

Diagrama de Trajetórias	**Variável Endógena**	**=**	**Variáveis Exógenas**	**+**	**Variáveis Endógenas**	**+**	**Erro**
	Y_1	=	$X_1 \quad X_2 \quad X_3$	+	$Y_1 \quad Y_2 \quad Y_3$	+	ε_i
Figura 21.(a)	Y_1	=	$b_1X_1 + b_2X_2$			+	ε_1
Figura 21.(b)	Y_1	=	$b_1X_1 + b_2X_2$			+	ε_1
	Y_2	=	b_3X_2	+	b_4Y_1	+	ε_2
Figura 22.(c)	Y_1	=	$b_1X_1 + b_2X_2$			+	ε_1
	Y_2	=	$b_3X_2 + b_4X_3$	+	$b_5Y_1 + b_6Y_3$	+	ε_2
	Y_3	=		+	$b_7Y_1 + b_8Y_2$	+	ε_3

Fonte: HAIR, 1998.

A escolha do tipo de matriz de entrada e estimação do modelo proposto

Em modelos de AT decidir sobre as formas de entrada de dados e o método de estimação do modelo geram impacto no resultado final. Modelos de AT focam no padrão das relações entre as variáveis e não em observações individuais, portanto, se usa apenas a matriz de variância e covariância ou de correlação como entrada de dados. Indica-se o uso de correlações quando pretende compreender o padrão de relações entre as variáveis, mas não explicar a variância total. Covariâncias são recomendadas quando se pretende testar uma teoria e explicar a variância total. Esse tipo de modelo é mais sensível às características da distribuição dos dados, principalmente no que se refere a normalidade ou forte curtose. A falta de normalidade pode inflacionar a estatística qui-quadrado e criar um viés ascendente em valores importantes para calcular a significância dos dados (HAIR et al., 1998). O uso de variáveis dicotômicas ou categóricas ordinais, comum principalmente na área da saúde, foco deste estudo, contribui para esse agravante (GARSON, 2015).

Embora não haja um consenso ainda com relação ao tamanho ideal da amostra para modelos AT, recomenda-se tamanho entre 200 e 400 para modelos com 10 a 15 variáveis afim de evitar estimativas de parâmetro instáveis e testes significância sem força (GARSON, 2015), emfim AT / MEE é considerada uma técnica para grandes amostras (HAIR, 1998).

Com relação a estimativa do modelo, o procedimento mais comum em software para calcular AT/MEE é a máxima verossimilhança (*Maximum Likelihood Estimation* - MLE) por

apresentar resultados eficientes e não viesados quando se atende a suposição de normalidade multivariada. No entanto, como nem todos os dados atendem aos requisitos de normalidade criou-se técnicas de estimação alternativas como mínimos quadrados ponderados (*Weighted Least Squares* - *W*LS), mínimos quadrados generalizados (*Generalized Least Squares* - GLS) e estimação assintoticamente livre de distribuição (*Asymptotically Distribution-Free* - ADF) (HAIR et al, 1998).

Avaliar a identificação do modelo

Analisar um modelo de AT não é trivial, nesta fase de identificação deve ser teoricamente possível obter uma estimativa única para cada parâmetro do modelo, caso contrário o modelo não poderá ser identificado. A palavra "teoricamente" enfatiza a etapa de identificação como uma característica própria do modelo e não dos dados. Portanto, se não for possível identificar um modelo, este permanecerá dessa forma, independente do tamanho da amostra (N=100, N=1.000, N=1.000.000). Por exemplo, um dos problemas de identificação é a falta de capacidade do modelo proposto para gerar as estimativas e a medida que o modelo se torna complexo mais difícil se torna o processo de identificação (HAIR. et al., 1998). Neste contexto, modelos não identificados devem ser especificados novamente e ser submetidos a etapa 1 novamente (FIGURA 22), pois tentativas de análises sobre este modelo serão em vão. Em se tratando de identificação do modelo, um termo comum é "graus de liberdade" (GL) que consiste na diferença entre o número de correlações ou covariâncias e o número real de coeficientes no modelo proposto. O cálculo deste termo se dá pela seguinte equação (Eq. 31):

$$\text{gl} - \frac{1}{2}[(p+q)(p+q+1)] - t \qquad (31)$$

Onde: p = número de indicadores endógenos, q = número de indicadores exógenos e t = número de coeficientes estimados no modelo proposto.

Os diversos softwares para AT / MEE possuem rotinas para encontrar possíveis problemas de identificação. Outros possíveis sintomas para problemas de identificação são: a) erros padrão de coeficientes muito grandes. b) incapacidade do software em inverter a matriz de informação, c) apresentação de estimativas exageradas ou irreais (variâncias negativas de erro) e d) correlações muito altas entre os coeficientes estimados. Ao se deparar com

problemas de identificação a solução é reduzir o modelo eliminando-se trajetórias e coeficientes (HAIR et al., 1998).

Avaliar estimativas do modelo e interpretá-lo

Essa etapa consiste em avaliar a "bondade de ajuste" do modelo e envolve o uso de uma ferramenta computacional capaz de gerar um modelo de AT e executar os procedimentos de análise necessários. Como um primeiro passo deve-se avaliar o ajuste do modelo, ou seja, determinar o quão bem o modelo explica os dados. Nesta etapa, existem duas possibilidades: a) o modelo não se ajusta bem aos dados, neste caso este deve ser reespecificado e b) o modelo apresenta um ajuste satisfatório, então deve-se prosseguir interpretando-se as estimativas dos parâmetros (KLINE, 2015).

Hair et al. (1998) ainda recomenda que se faça uma inspeção inicial em busca de "estimativas transgressoras", ou seja, verificar coeficientes que excedam os limites aceitáveis como variâncias negativas ou sem significância de erros para variáveis e coeficientes padronizados com valores excedentes ou muito próximos de 1,0. Caso o modelo seja aceitável, avalia-se o ajuste geral do modelo por meio das medidas de bondade de ajuste. Avaliar se o modelo se ajusta bem ao dados significa que a matriz de covariância ou correlação prevista pelo modelo proposto está bem próxima da matriz dos dados de entrada observados. O resultado gerados por modelos de AT podem ser seriamente afetados por multicolinearidade, portanto recomenda-se ficar avaliar valores de correlação superiores a 0.90 e observar valoresacima de 0,80.

Kline (2015) recomenda também considerar modelos equivalentes ou semi-equivalentes, pois um modelo equivalente que apresenta uma configuração diferente de relações hipotéticas entre as mesmas variáveis, pode explicar os dados tão bem quanto o modelo inicial. Neste caso, fica a critério do pesquisador decidir qual modelo utilizar.

Nesta fase o pesquisador deve avaliar o modelo e identificar se este é aceitável ou não. Aceitável se refere ao modelo com bom ajuste ao dados e neste caso o pesquisador deve interpretar as estimativas dos parâmetros e se for o caso considerar modelos equivalentes ou semi-equivalentes. Com o modelo final em mãos, este deve ser descrito detalhadamente e os resultados de análise apresentados (HAIR et al., 1998; KLINE, 2015).

Reespecificar o modelo

Se o pesquisador identificar que o ajuste do modelo é ruim, possivelmente devido a sua complexidade ou por possuir muitas restrições este deve retornar a lista de possíveis modificações teoricamente justficáveis a qual se referiu em seção anterior e refazer o modelo visando melhorar o modelo ou até mesmo simplificá-lo. Este passo deve ser executado de modo que a reespecificação seja guiada mais por considerações racionais do que estatísticas (KLINE, 2015).

Nota final

Embora Hair et al. (1998) não tenha citado em seus passos, Kline (2015) recomenda uma fase de "triagem dos dados". Esses passos visam a busca por colinearidades, onde variáveis separadas medem a mesma coisa e neste caso uma ou outra deve ser incluída no modelo, não ambas. Deve-se buscar por *outliers* (pontuações muito diferente das demais) e dados ausentes. E finalmente verificar se o pressuposto de normalidade pode ser atendido ou nao, caso contrário utilizar métodos de estimação apropriado como citado em seções anteriores.

2.5.7 Pressupostos

Por ser uma extensão da regressão linear, AT assume vários dos pressuposto desta técnica. Primeiro, as relações entre as variáveis devem ser lineares. Segundo, não deve ocorrer interação entre variáveis(embora possamos adicionar um novo termo que reflita a interação de 2 variáveis). Terceiro, variáveisendógenas devem ser contínuas (embora você possa fugir com um mínimo de 5 categorias, se você tiver dados ordinais) e relativamente distribuído, com coeficientes de aspereza e curtose abaixo de 1. Quarto, presume-se que as covariâncias entre os termos de perturbação sejam zero (equivalente à hipótese de erros não correlacionados entre as variáveispreditoras na regressão). Versões mais avançadas de AT podem lidar com violações desta suposição. Por fim. AT é muito sensível à especificação do modelo, incluir variáveis irrelevantes ou omitir as relevantes, podem afetar drasticamente os resultados.

2.6 CONSIDERAÇÕES FINAIS

Este capítulo apresentou uma revisão bibliográfica com foco nos principais temas abordados nesta tese, que são a inferência causal estatística, modelos gráficos, redes Bayesianas e Análise de Trajetórias. Com relação a inferência causal, foi feita uma abordagem acerca da importância com relação análise de causalidade e os cuidados que devem se tomar ao se criar esse tipo de modelo. Conceitos importantes sobre a criação de modelos gráficos foram apresentados, pois para os assuntos que se seguiram se considerou importante conhecer estes conceitos. Também foram explorados os principais conceitos sobre RBs, bem como apresentados os algoritmos de aprendizagem de estrutura a partir de dados. Por fim se dedicou também um seção definir a análise de trajetórias e seus principais conceitos, bem como formas de construir e avaliar a qualidade dos modelos de AT.

O próximo capítulo apresenta estudos correlatos, os quais utilizaram a criação de modelos de RBs, ATs e MEES como solução para problemas da área da saúde.

3 ESTUDOS CORRELATOS

3.1. CONSIDERAÇÕES INICIAIS

Este capítulo trata sobre os trabalhos relacionados à RBs e ATs encontrados na literatura. Por esta tese ter objetivo de executar experimentos na área pesquisa clínica, especialmente cardiologia, a busca na literatura foi por estudos nesta mesma área. Além disso, incluiu estudos envolvendo MEE, pois AT tem sido muito utilizado via MEE em estudos das áreas de economia, sociologia e principalmente em ciências comportamentais, mas pouco utilizado na área da pesquisa clínica. A Seção 4.1 apresenta estudos que utilizaram RBs em estudos sobre doenças cardiovasculares. Na Seção 4.2 são descritos os estudos envolvendo ATs e MEEs na mesma área. Na Seção 4.3 são feitas algumas considerações sobre os trabalhos citados.

3.2 A UTILIZAÇÃO DE REDES BAYESIANAS NA ÁREA DA SAÚDE

Considerando as RBs, Al-Hamadani (2016) propôs um sistema especialista com o objetivo de ajudar médicos de unidades de emergência em diagnosticar doença que provoca insuficiência cardíaca. Este sistema utiliza RBs para modelar a estrutura da rede com nós discretos a partir de informações incertas ou incompletas e utiliza regras de linguagem de semântica para construir regras de inferência e *Java Expert System Shell* (JESS) como mecanismo de inferência. O modelo atingiu 75% de acurácia e 83% de sensibilidade e 66% de especificidade. Oliveira, Andreao e Sarcinelli Filho (2016) sugeriram uma Rede Bayesiana Dinâmica (RBD) para melhorar as decisões médicas quando se lida com o problema da classificação de batimentos em eletrocardiogramas. Essa RBD considera a incerteza toda vez que aparecem novas evidências durante o processo de tomada de decisão. A RBD alcançou sensibilidade e previsão positiva de 99% para batimentos ventriculares prematuros demonstrando que RBF é uma ferramenta promissora para classificar arritmias cardíacas. Orphanou, Stassopoulou e Keravnou (2016) desenvolveram um modelo de RBD extendida que integra métodos de abstração temporal (AT) como uma ferramenta para sistemas de apoio à decisão médica e geração de um modelo prognóstico para risco de doença coronariana. A estrutura da rede foi construída por meio da derivação de ATs e algoritmos de aprendizagem de máquinas para aprender os parâmetros do modelo. A curva ROC deste modelo contra um modelo de RND sem TA atingiu 78% de área sob a curva *Receiver Operating Characteristic*

(ROC). Wei et al. (2016) utilizaram uma base de 10.792 casos e construíram um modelo de RB usando o algoritmo *tabu search n*para construir a estrutura e a máxima verossimilhança para calcular a TPC de cada nó. Os resultados demonstraram que o modelo foi capaz de revelar as correlações complexas entre os fatores de influência na DCV e a relação com as doenças coronarianas. Gatti, Luciani e Stella(2012) utilizaram redes Bayesianas de tempo contínuo (RBTC) para identificar insuficiência cardíaca cardiogênica aguda e antecipar a sua provável evolução por meio do desenvolvimento de um plano estratégico para reduzir o risco associado ao tratamento de cada paciente. A validação do componente qualitativo do modelo (estrutura gráfica) se deu por comparação entre o modelo gerado e exemplos de um livro texto de cardiologia. E do componente quantitativo foi por meio de perícia médica. Os resultados se mostraram consistentes com a atual compreensão médica patofisiológica de quadros clínicos da doença. Flores et al. (2001) utilizaram dados de domínio público para insuficiência cardíaca para desenvolver um sistema de detecção causal automatizado. Este sistema permite a incorporação de vários tipos de conhecimentos especializados prévios para testar e comparar descobertas não tendenciosas com a descoberta enviesada e com diferentes tipos de opinião de especialistas. Foram utilizadas matrizes de adjacência aprimoradas com rótulos numéricos e coloridos para auxiliar na interpretação dos resultados. Para o problema estudado os resultados se mostraram mais eficazes em ajudar na descoberta de modelos do que usar informações anteriores ou antecedentes especializados detalhados.

3.3 A UTILIZAÇÃO DE ATS NA ÁREA DA SAÚDE

Singh et al.(2016) propuseram uma nova abordagem para modelar DCV que combinou MEE/AT e Mapa Cognitivo *Fuzzy* (MCF) para diagnosticar doenças cardíacas. O conjunto de dados para execução dos experimentos foi o CCHS. O qui-quadrado foi utilizado com variáveis categóricas com mais de 6 categorias e o p-valor entre cada variável e a variável que indica se o paciente tem DCV foi utilizado. Os relacionamentos bivariados que apresentaram p-valor mais significante foram classificados em ordem crescente e a partir dos valores mais significantes foram selecionadas 20 variáveis. Essas variáveis foram agrupadas em 6 categorias que foram utilizadas como variáveis latentes. A partir desses relacionamentos se criou o modelo de MEE/AT. As arestas desse modelo foram utilizadas para criar uma matriz de pesos representando a força do relacionamento causal entre as variáveis. Essa matriz de pesos foi utilizada para criar o Mapa Cognitivo *Fuzzy* (MCF), o modelo gerado atingiu 79% de área sob a curva ROC (*Receiver Operating Characteristic*) e 74% de precisão. Chen,

Srinivasan e Berenson (2008) utilizarammodelos de AT para dissecar as relações complexas de índice de massa corporal (IMC) e insulina em jejum com outros componentes da síndrome metabólica (SM) e também as diferenças entre brancos e negros nessas relações. O modelo de AT foi criado com base em achados anteriores e raciocínios teóricos. Os parâmetros do modelo foram estimados por meio do método de máxima verossimilhança. Os índices de ajuste do modelo variaram de 0,927 a 0,985 indicando um bom ajuste dos seis modelos criados aos dados. Ao final o estudo demonstrou que a obesidade infantil é o fator mais crítico que contribui para o desenvolvimento da síndrome metabólica. Kim (2007) criou um modelo de AT para avaliar a qualidade de vida relacionada a saúde de pacientes com insuficiência cardíaca a partir de um conjunto de dados com 103 pacientes. O modelo foi criado a partir de variáveis derivadas de estudos anteriores provenientes do modelo hipotético de Wilson e Cleary. Os índices de ajuste do modelo GFI, AGFI, NNFI, NDI e p-valor foram utilizados para avaliar o ajuste geral do modelo e todos apresentaram valores que indicaram um bom ajuste do modelo aos dados. Ao final conclui-se que para melhorar a qualidade de vida desses pacientes é necessário fazer intervenções de enfermagem para o melhorar o funcionamento físico e diminuir a depressão.

Também identificou-se modelos que utilizam MEE que é a grande área de AT. Castro et al. (2015) avaliaram a inter-relação entre os indicadores de obesidade geral e central como preditores de doenças cardiovasculares metabólicas usando modelos de MEE. O model avaliou as inter-relações entre as índice de massa corporal (IMC), circunferência da cintura (CC), inflamação, pressão arterial e perfil lipídico. Os resultados mostraram que ambos atuam como preditores diferentes, e a relação cintura / altura superou o índice de massa corporal e a circunferência da cintura como preditores. Não foram relatados os índices do modelo de MEE criado. Vellone et al.(2013) desenvolveram um MEE para avaliar a situação de autocuidado de pacientes com falha cardíaca. O objetivo do estudo foi de melhorar o conhecimentodo processo e das relações entre os conceitos teóricos de aderência ao tratamento e a avaliação do tratamento. O modelo hipotético deste estudo foi criado por MEE. O ajuste do modelo foi avaliado considerando-se qui-quadrado, CFI, RMSEA e SRMR. Inicialmente o modelo apresentou índices insatisfatórios e por isso foi reespecificado. O resultado final mostrou que o monitoramento de sintomas tem um relacionamento direto e positivo com a implementação do tratamento e que a avaliação do reconhecimento dos sintomas tem uma relação direta e positiva com avaliação do tratamento. De Heer et al.(2008) investigou a efetividade de uma intervenção educacional na comunidade promovida por agentes de saúde para reduzir os riscos de DCV entre pacientes espanhóis por meio de um modelo de MEE. O modelo de MEE

foi criado a partir de um modelo conceitual publicado por Anders, Balcázar e Paez(2006) de prevenção e redução de doenças cardiovasculares entre mexicanos-americanos. O modelo foi estimado por meio da máxima verosimilhança e a avaliação do modelo foi executada por meio do qui-quadrado, RMSEA, NNFI, CFI. Todos os índices apresentados pelo modelo comprovaram que este obteve um ajuste aceitável.

3.4 CONSIDERAÇÕE FINAIS

Na maioria dos artigos, os dois métodos foram usados de forma independente para construir modelos preditivos e ajudar na prevenção de DCV. Embora alguns artigos não tenham especificados, observou-se que os modelos foram construídos com a ajuda de especialistas, provavelmente estatísticos e especialistas em RBs, AT e / ou MEE. Sabe-se que esse profissionais podem ser caros e demorar para apresentar o modelo final. Nenhum dos estudos investigados usou algoritmos para aprender modelos preditivos a partir de dados conforme propõe o método proposto por este estudo. De acordo com o conhecimento do autor desta tese, este estudo é único a propor a construção de modelos AT a partir de um conjunto de dados usando AAERBs.

O próximo capítulo apresenta de forma detalhada o método proposto para o desenvolvimento, execução de experimentose avaliação de resultados do método *bnpa* como solução pretendida para o problema apresentado.

4 MÉTODO PROPOSTO

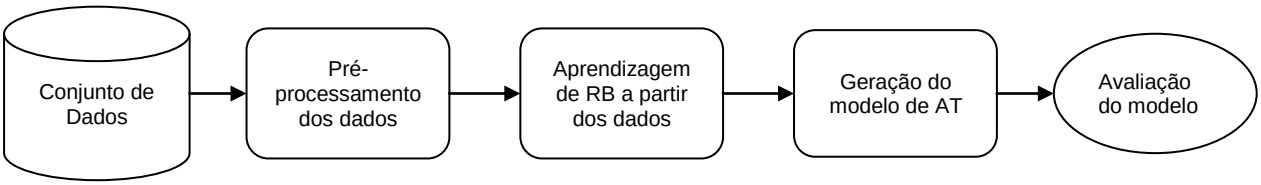
4.1 CONSIDERAÇÕES INICIAIS

Neste capítulo serão apresentados os componentes necessários para o desenvolvimento desta tese de doutorado. Neste contexto a seção 4.2 apresenta-se em detalhes o método proposto, destacando-se cada uma das etapas que o compõe. Na seção 4.3 apresenta-se os detalhes referentes a como os experimentos serão realizados. Finalmente na seção 4.4 conclui-se o capítulo.

4.2 O PROCESSO SEMI-AUTOMÁTICO DE CRIAÇÃO DE MODELOS DE ANÁLISE DE TRILHAS POR MEIO DE APRENDIZAGEM DE RBS

Para a realização dos experimentos desta tese foi necessário a execução de 4 etapas, sendo elas: pré-processamento por meio do qual dados problemáticos devem ser identificados e tratados, aprendizagem da estrutura da RB para que se obtenha uma estrutura gráfica causal, geração do grafo de AT que permitirá a avaliação de efeitos diretos e indiretos das variáveis, avaliação do modelo e interpretação dos resultados obtidos. Esse processo é ilustrado pela Figura25. Em seguida cada uma dessas etapas serão detalhadas.

Figura 25 - Fluxo de tarefas a serem executadas pelo método proposto



Fonte: o autor, 2018.

4.2.1 Pré-processamento dos conjuntos de dados

Dados faltantes,pontos de dados influentes (*outliers)*, multicolinearidade e não normalidade são características comumente encontradas em conjuntos de dados que refletem a realidade. Todos esses são fatores que podem afetar seriamente o processo de estimação de modelos de AT, pois podem contribuir com a geração de variâncias negativas, matrizes definidas não positivas e falhas em alcançar a convergência (incapacidade de calcular um conjunto de estimativas de parâmetros) (SCHUMACKER; LOMAX, 2012). Frente a esses

possíveis problemas, a criação do método proposto por este estudo resultou na criação de diversas pequenas funções. Entre essas funções, existem três tipos: as que auxiliam o pré-processamento, auxiliam na criação da RB e na criação do modelo AT. No entanto, algumas, embora não menos importantes, são genéricas. Como por exemplo a função que identifica os tipos de variáveis que compõem o conjunto de dados. Quando executada, esta função indica se a variável é quantitativa contínua ou discreta, se é qualitativa dicotômica, ordinal ou nominal, informação de extrema importância para nortear os procedimentos a serem executados. Por exemplo, durante a aprendizagem da estrutura de RBs é necessário saber o tipo de variável existente no banco de dados que será utilizado em conjunto com o método *bnpa*. Neste caso, é preciso descobrir se a variável é contínua, categórica dicotômica, categórica ordinal ou categórica nominal para então decidir quais testes usar em conjunto com o algoritmo de AAERBs. Por exemplo, se o algoritmo a ser processado for o '*gs*' que é baseado em restrições verifica-se o tipo de variável. Se a variável for categórica ordinal o teste de independência condicional a ser executado deve ser o 'Joncheere-Terpstra' ou 'jt', se o tipo for categórica nominal ou dicotômica os testes disponíveis são '*mutual information*' ou 'mi', '*shrinkage estimator for the mutual information*' ou 'mi-sh' e 'qui-quadrado' ou 'x2'. Já no caso do algoritmo '*hc*' que é baseado em pontuação e no caso de variáveis categóricas o método de pontuação pode ser "*multinomial log-likelihood*" ou "loglik", "*Akaike information*" ou "aic" entre outros. No caso de algoritmos baseados e pontuação independe se a variável é categórica dicotômica, ordinal ou nominal, pois todas são tratadas da mesma maneira. Portanto, para o método *bnpa* a função que identifica o tipo de variável no banco de dados a ser trabalhado e de vital importância, uma vez que o processo para aprender a RB é automático.

A seguir serão descritos algumas ações executadas no método *bnpa* como métodos auxiliares a serem executados quando necessários. Todo esse processo de pré-processamento exige atenção especial do pesquisador, foram criadas funções auxiliares que estão embutidas no método *bnpa* as quais devem ser utilizadas de forma manual pelo pesquisador. Após o pré-processamento todo o processo será automático, gerará grafos, tabelas e textos informativos para análise pelo pesquisador. No entanto após a criação automática das estruturas de RBs e antes de iniciar o processo automático de criação do grafo de AT, especialistas devem avaliar se os modelos de RBs aprendidos servem para continuar o processo ou devem ser reparametrizados. Maiores detalhes sobre o papel desses especialistas serão descritos nas próximas seções.

Remoção, recodificação e transformação de valores

Respostas que não eram significantes para este estudo como por exemplo: “Não aplicado”, “Não sabem”, “Recusa” e “Nada a declarar” foram eliminadas. Variáveis dicotômicas com respostas como “Não” e “Sim” foram recodificadas para “0” e “1” respectivamente. Dados codificados de forma dicotômica ou ordinal foram transformados em variáveis categóricas.

Tratamento de dados faltantes

Ao processar dados epidemiológicos é comum se deparar com dados faltantes, por exemplo, nem todos os pacientes têm dados sobre gestação. Portanto, por este estudo ser voltado para a pesquisa epidemiológica se considerou tratar esse tipo de problema.

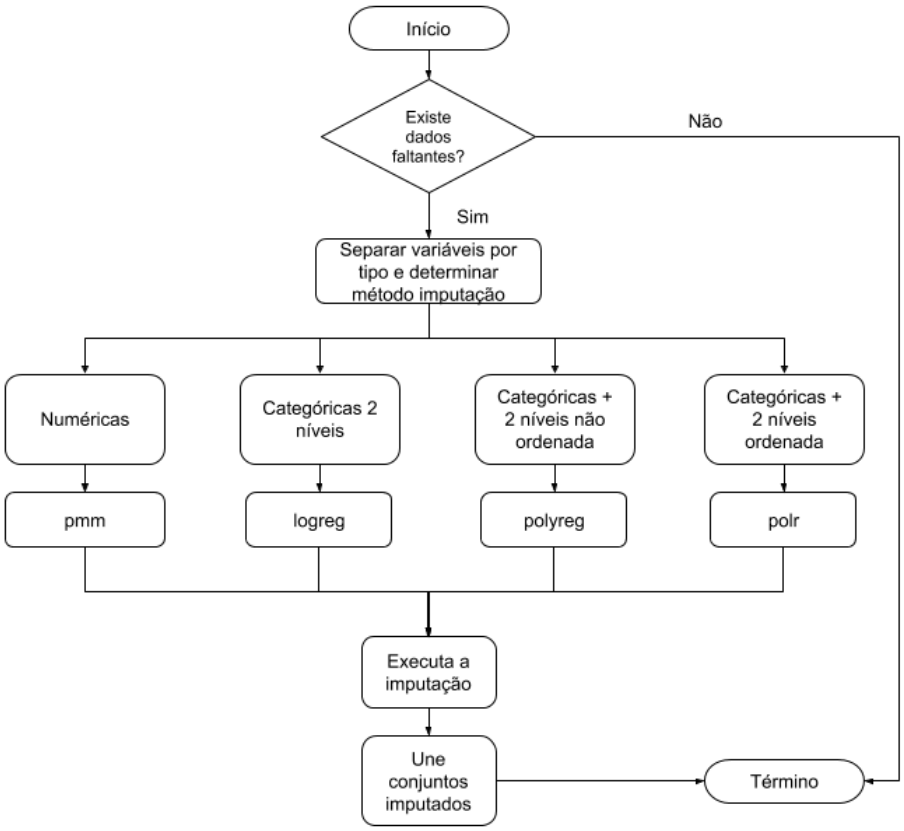
É normal que conjunto de dados possuam dados faltantes, portanto estes devem ser identificados e um percentual de quanto representam para a amostra deve ser identificado. Grandes percentuais de dados faltantes (*missings*) requerem avaliação do pesquisador quanto a remoção ou método de imputação a ser executado sobre esses dados. Os critérios devem ser baseados na área de pesquisa e com base na literatura encontrada.

Para este estudo utilizou-se como apoio o pacote R MICE (VAN BUUREN, 2015), um *software* para ambiente R que proporciona imputação múltipla por meio de Especificação Totalmente Condicional (ETC) (*Fully Conditional Specification* - FCS) implementado pelo algoritmo MICE. Este pacote de *software* foi escolhido pelo fato de o mesmo fornecer modelos de imputação para dados contínuos (correspondência de média preditiva, normal - *predictive mean matching*), dados binários (regressão logística), dados categóricos não ordinais (regressão logística politômica) e dados categóricos ordinais (odds proporcionais).

A ideia do método proposto por este estudo é automatizar o máximo possível as tarefas a serem realizadas pelo pesquisador. Portanto, com relação aos dados faltantes criou-se uma implementação de *software* (*check.na*) para identificar dados faltantes e executar a imputação múltipla (*input.missing.data*) usando o pacote MICE. A Figura 26 mostra o fluxo de execução para imputação de dados faltantes embutida no método *bnpa*. Na primeira etapa do processo se confirma a real existência de dados faltantes. Em seguida, separa-se as variáveis por tipo numérica, categórica de dois níveis, categórica com mais de dois níveis não ordenada e categórica com mais de dois níveis ordenada. Essa separação se faz necessário para que ao executar o processo de imputação múltipla seja empregado o método correto, ou

seja, pmm (*predictive mean matching*) para variáveis numéricas, logreg (*logistic regression*) para variáveis categóricas de dois níveis, polyreg (*polytomous regression*) para variáveis com mais de dois níveis não ordenadas e polr (*proportional odds model*) para variáveis categóricas com mais de dois níveis ordenadas. Em seguida executa-se a imputação criando-se por default 10 imputações e se une o conjunto de dados.

Figura 26- Fluxograma ilustrando o processo de imputação múltipla



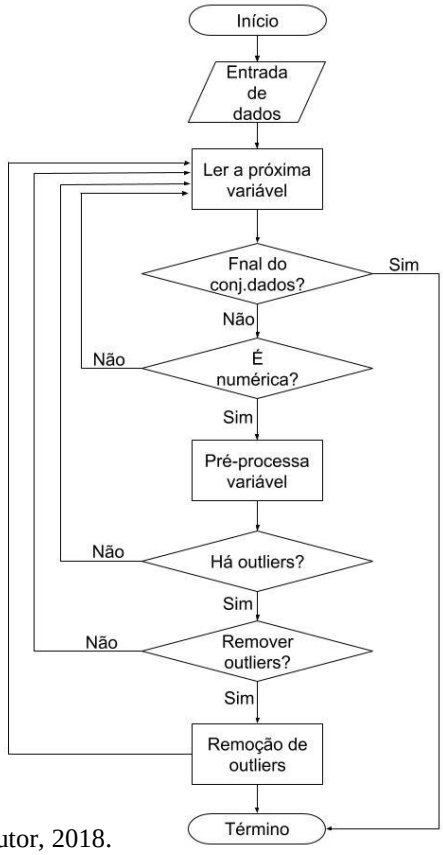
Fonte: o autor, 2018.

Identificação de pontos de dados influentes (*outliers*)

Define-se *outliers* ou pontos de dados influentes como valores de dados extremos ou atípicos nas VIs ou VDs ou em ambas. Esses resultados podem ocorrer como resultados de erros de observação, erros de entrada de dados, erros de instrumento baseados em *layouts* ou instruções ou valores extremos reais de dados. Pontos de dados influentes normalmente refletem no resultado da média, do desvio padrão e nos valores de coeficiente correlação. Esses valores extremos devem sempre ser identificados, explicados, excluídos ou acomodados usando-se métodos estatísticos disponíveis. E em algumas situações dados adicionais precisarão ser coletados para preencher a lacuna ao longo dos dados (SCHUMACKER; LOMAX, 2012). Pelo fato de dados epidemiológicos serem muitas vezes coletados de forma

manual e conter erros não é incomum encontrar *outliers* entre eles, portanto, também se optou por implementar uma solução para tratar este problema. O fluxo de processos para identificar *outliers* está ilustrado pela Figura 27 e foi implementado no método bnpa *(check.outliers*). Este processo se inicia com a submissão do conjunto de dados como entrada de dados, em seguidase faz uma varredura de cada variável do mesmo, durante a leitura da variável se verifica a mesma é numérica e se executa um pré-processamento. Durante durante o pré-processamento da variável é gerado um diagrama de caixas (boxplot) por meio do pacote R grDevices (TEAM, 2017). A determinação de *outliers* deste pacote é executada por meio da extensão dos "bigodes" (*whiskers*), que é baseada no comprimento da "caixa" (entre os quartis inferior e superior) e o coeficiente multiplicador cujo valor é 1.5. Após a exposição dos *outliers* se apresenta a média e a distribuição dos dados por meio de um histograma. Em seguida uma nova média sem os *outliers* e um novo histograma mostrando a distribuição dos dados são apresentados, sendo que neste segundo gráfico a distribuição tende a estar normal. Após isso, se pergunta se o pesquisador deseja remover os *outliers* e caso se responda “Sim” eles serão removidos e uma nova iteração reiniciará até que todos os *outliers* de todas as variáveis sejam removidos.

Figura 27 - Fluxograma ilustrando o processo de identificação e remoção de outliers

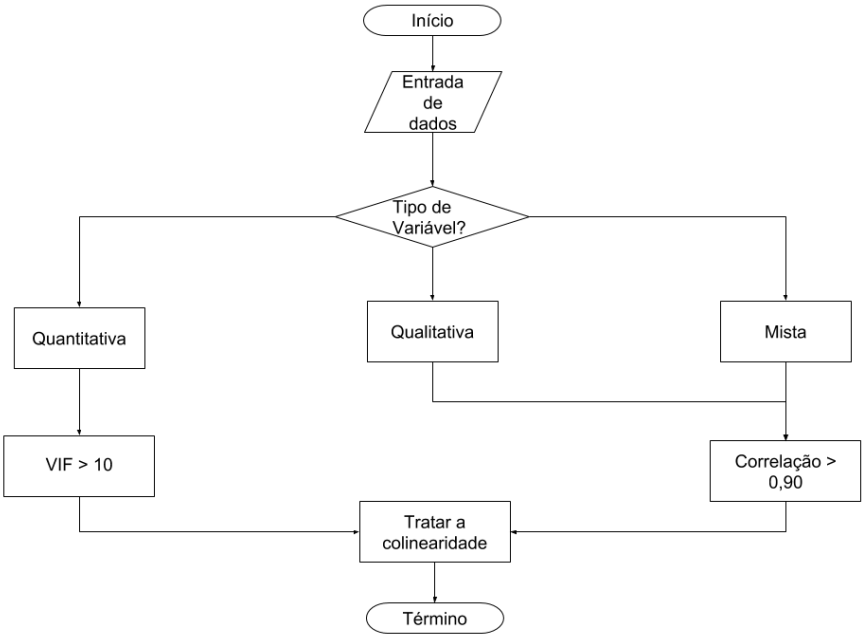


Fonte: o autor, 2018.

Identificação e multicolinearidade

Os coeficientes de um grafo de AT podem ser inflacionados na ocorrência de multicolinearidade no conjunto de dados, seja esta severa ou moderada. Neste caso, os coeficientes por assumirem valores demasiadamente altos resultam em estimativas não confiáveis. Dentro deste cenário, se faz necessário adotar sempre um procedimento para diagnosticar a multicolinearidade antes de iniciar este tipo de modelagem (GUJARATI; PORTER, 2011). Portanto, para verificar a existência de colinearidade, se adotou o procedimento ilustrado pela Figura 28, onde o primeiro passo é verificar o tipo de variável que o conjunto de dados possui. Caso as variáveis sejam identificadas como quantitativas, será executada um procedimento implementado no método *bnpa* (*check.collinearity*) para calcular o Fator de Inflação de Variação (VIF) e aqueles casos em que os valores de VIF sejam maiores que 10 serão sugeridos como casos que geram colinearidade. No caso de variáveis identificadas como quantitativasou mistas (quantitativa e qualitativas) será calculado e examinado o coeficiente de correlação. Neste caso será utilizado o pacote R polycor (FOX, 2010), uma vez que este calcula correlação polisserial entre variáveis numéricas e ordinais e correlação policórica entre variáveis ordinais.Considerando esta situação as variáveis que apresentarem o correlação bivariada acima de 0,90 serão consideradas como possíveis causas de multicolinearidade.

Figura 28 - Fluxograma ilustrando o processo de verificação da colinearidade



Fonte: o autor, 2018.

Após o processo de identificação de possíveis variáveis como causa de colinearidade, o método as apresenta para o pesquisador e este deve pensar o que fazer: ignorar ou remover algumas variáveis.

A próxima seção apresenta qual foi o processo adotado para a construção da RB para este estudo. Este processo adotou o pacote de *software* para o ambiente estatístico *R* chamado bnlearn.

4.2.2 Construção da estrutura da rede Bayesiana

O processo de construção de uma RB requer duas etapas: a primeira onde se constrói a estrutura do grafo de RB e a segunda onde se estima a TPC (KORB; NICHOLSON, 2003). Neste contexto, a estrutura da RB pode ser concebida por meio da ajuda de especialistas ou ser aprendida a partir de um conjunto de dados por meio de algoritmos de aprendizagem de estrutura de RBs conforme descrito na seção 2.4.3. Para este estudo, se escolheu a segunda opção, pois o propósito do método proposto por esta tese é fornecer autonomia para que o pesquisador possa criar seus modelos de análise causal, inicialmente sem a ajuda de um estatístico ou modelador causal.

O uso do pacote bnlearn como *software* de apoio

Como o objetivo deste estudo é capacitar o método de AT a identificar causalidades para criar modelos a partir de dados, buscou-se por métodos que tivessem essa capacidade e estivessem consolidados na literatura. Entre os métodos pesquisados, o método para criação de RBs foi o que melhor se adequou ao objetivo proposto, pois além de possuir diversos algoritmos para aprendizagem de estrutura de RBs a partir de dados, necessidade deste estudo, também haviam diversas implementações no ambiente R, o mesmo ambiente em que o método *bnpa* foi implementado.

Como *software* de apoio para a construção de RBs, se escolheu o pacote R bnlearn (SCUTARI, 2014). O primeiro motivo foi por este estar implementado no ambiente estatístico R. O segundo motivo foi implementar os seguinte algoritmos para aprendizado de estruturas de RBs (AAERB): a) algoritmos baseados em restrição: PC, *Grow-Shrink* (GS), *Incremental Association Markov Blanket* (IAMB), *Fast Incremental Association* (Fast-IAMB), *Interleaved Incremental Association* (Inter-IAMB), *Max-Min Parent Children* (MMPC), *Semi-Interleaved Hiton-PC* (SI-HITON-PC); b) algoritmos baseados em pontuação: *Hill Climbing* (HC), *Tabu*

Search (Tabu) e c) algoritmos híbridos:*Max-Min Hill Climbing* (MMHC), *General 2-Phase Restricted Maximization* (RSMAX2). O terceiro motivo é que o bnlearn suporta dados discretos (multinomiais) e contínuos (normais multivariados) para os processos de aprendizagem de estrutura de RBs. O quarto motivo é que este pacote de *software* para cada algoritmo de aprendizagem oferece ainda diversos testes de independência condicionais de acordo com o tipo de variável, conforme se apresenta a seguir.

Para os algoritmos de aprendizagem de estrutura baseados em restrições, os testes de independência implementados no bnlearn são (SCUTARI, 2014):

1. dados categóricos (distribuição multinomial):
   - informação mútua;
   - estimador de encolhimento para informação mútua;
   - Qui-quadrado de Pearson.

2. dados ordinais:
   - *Jonckheere-Terpstra*.

3. dados contínuos (distribuição normal e multivariada):
   - correlação linear (*linear correlatio*n);
   - teste Z de Fisher (Fisher's *Z*);
   - informação mútua;
   - estimador de encolhimento (*shrinkage-estimator*) para informação mútua.

4. dados mistos (distribuição condicional Gaussiana):
   - informação mútua (mutual information).

Para os algoritmos de aprendizagem de estrutura baseados em pontuação, os testes de pontuação implementados no bnlearn são (SCUTARI, 2014):

1. dados categóricos (distribuição multinomial):
   - log de probabilidade multinomial;
   - critério de informação Akaike;
   - critério de informação Bayesiana;
   - uma pontuação equivalente à densidade posterior de Dirichlet;

- densidade esparsa posterior de Dirichlet;
- uma densidade posterior de Dirichlet baseada no anterior de Jeffrey;
- um Dirichlet bayesiano modificado para dados mistos intervencionais e observacionais ;
- pontuação BDe de média local;
- pontuação K2.

2. dados contínuos (distribuição normal e multivariada):
   - a probabilidade de log multivariada gaussiana;
   - o correspondente Critério de Informação de Akaike;
   - o correspondente Critério de Informação Bayesiano;
   - uma pontuação equivalente densidade posterior Gauss (BGe).

3. dados mistos (distribuição condicional Gaussiana):
   - a probabilidade verossimilhança condicional de Gauss;
   - o correspondente Critério de Informação de Akaike;
   - o correspondente Critério de Informação Bayesiano.

A maioria dessas funcionalidades serão automaticamente selecionadas pelo método *bnpa* durante o processo de aprendizagem de estrutura de RB de acordo com o algoritmo de AAERBs selecionado e do tipo de variável identificada. Por isso a importância de indentificar o tipo de variável conforme descrito na seção 4.2.1.

Mais um motivo para utilizar o pacote bnlearn (SCUTARI, 2014) é o fato deste ter implementado a validação cruzada, uma maneira padrão de obter estimativas imparciais da qualidade do ajuste de um modelo. Ao comparar estas estimativas para diferentes algoritmos de aprendizagem de estrutura de RBs é possível escolher a melhor estrutura de RB aprendida, que idealmente seria e que apresentar a menor taxa de erro.

A dificuldade de aprender estruturas de RBs para estudos clínicos

Durante a implementação do método e realização dos testes iniciais, percebeu-se o seguinte problema: quando se faz modelagem por meio de RBs, no DAG resultante não existem variáveis determinadas especificamente como causa de outras ou vice-versa. Isso significa que qualquer variável pode apontar para qualquer outra. Por outro lado, em estudos

epidemiológicos onde se analisa estatisticamente a relação entre variáveis, é frequente a presença de VDs e VIs, ou seja, nem toda variável pode apontar para qualquer outra e vice-versa. Por exemplo, neste tipo de estudos, "idade" normalmente apontando para outras variáveis é razoável, mas quase nenhuma outra variável aponta para ela. Portanto, normalmente "idade" atua como um típico preditor (VI), ou seja, ela pode apontar para "pressão alta", "diabetes" e "doença cardiovascular", mas essas três variáveis não devem apontar para ela. É claro que exceções existem, por exemplo a variável "tempo" pode apontar para "idade" como sua causa. Desta maneira, cada estudo precisa ser avaliado no que se refere a como as variáveis serão modeladas e dispostas no modelo causal. Já para variáveis tipicamente de desfecho deve ocorrer o contrário, ou seja, ela não deve apontar para nenhuma outra, mas qualquer uma pode apontar para ela. Por exemplo "doença cardiovascular" não pode apontar para "idade" ou "fumar".

Para tratar este problema, se desenvolveu o algoritmo 1, cuja função é utilizar a função "lista negra" ou "*black-list*" do pacote bnlearn (SCUTARI, 2014) que cria pares de conexões que não serão permitidas durante a aprendizagem da estrutura da RB. Para maior esclarecimento sobre essa funcionalidade deve-se ter em mente que a primeira etapa para aprender uma RB é gerar a sua estrutura. A geração dessa estrutura pode ser feita a partir de dados, ou seja, os dados podem ser usados para determinar quais arcos estarão presentes no grafo sujacente ao modelo. Seria perfeito se esse processo fosse puramente orientado a dados, principalmente quando não se sabe muito sobre o fenômeno a ser modelado. No entanto, geralmente se tem conhecimento prévio sobre como deve ser construída a estrutura da RB e este conhecimento pode ser incoroporado no processo de aprendizado da estruutra da RB. Isso pode ser feito no caso do pacote bnlearn utilizando-se o parâmetro 'whitelist' cujos arcos são sempre incluídos na rede e/ou parâmetro 'blacklist' cujos arcos nunca são incluídos na rede (SCUTARI. 2014). Essa funcionalidade supriu as necessidades exigidas pelo algoritmo 1.

O algoritmo funciona da seguinte maneira: a) se passa como parâmetros o conjunto de dados, o nome da variável, o tipo de variável, sendo "o" para *outcome* (desfecho) ou "p" para *predictor* (preditor) e uma lista negra vazia ou pré-preenchida; b) Se executa uma varredura em todos os nomes de variáveis do conjunto de dados; c) Se verifica se o nome da variável sendo lida é o mesmo que a variável a ser processada e em caso positivo esta será ignorada; d) Se o tipo da variável passada for "*outcome*" se monta uma lista desta variável apontando para a variável atual; e) Se o tipo da variável for "*predictor*" se cria uma lista de todas as variáveis apontando para ela. Sendo um dos objetivos do método proposto por este estudo facilitar o trabalho de pesquisadores novatos, este algoritmo, além de automatizar a criação de listas

negras também converterá automaticamente para o formato do pacote bnlearn, cuja sintaxe não é trivial.

Figura 29 - Algoritmo 1 – Gerador de listas negras para o processo de aprendizagem de esturutras de RBs

Algoritmo 1: Gerador de listas para variáveis preditoras e de desfecho

Entrada: *d*, *v*, *t*, *l*

d = o conjunto de dados
v = o nome da variável a ser definica com preditor ou desfecho
t = tipo de variável (o = desfecho, p = preditor)
l = a lista negra (pode estar vazia ou pré-preenchida)

Saída: Uma lista negra

 # Inicia processo de montagem da lista de acordo com o tipo de variável

Para cada variável do conjunto de dados faça: # Percorre todas as variáveis do conjunto de dados
    Se o nome da variáveç atual <> de *v* então # Exclui a variável preditora ou de desfecho
        Se o tipo é desfecho
            l = l + v apontando para a variável atual
        else se o tipo é preditor
            l = l + o nome da variável atual apontando para *v*
        Fim se
    Fim se
Fim para

Fonte: o autor, 2018.

A sintaxe a ser utilizada pelo pesquisador é simples, pois ele precisa apenas chamar função responsável pelo processo passando como parâmetro o conjunto de dados, o tipo de variável, o nome da mesma. Tendo em mão a lista negra completa, se chama o processo que converte esta lista no formato para o pacote bnlearn (SCUTARI, 2014) processá-la (FIGURA30)

O resultado final deste algoritmo é uma lista com todos os pares de variáveis proibidas na estrutura de RB a ser aprendida. O formato desta lista aparece na parte superior da Figura 31 Em seguida esta lista será convertida para o formato exigido pelo pacote bnlearn apresentado na parte inferior da Figura 31.

Figura 30 - Sintaxe do método bnpa para determinar variáveis como desfechos ou preditoras

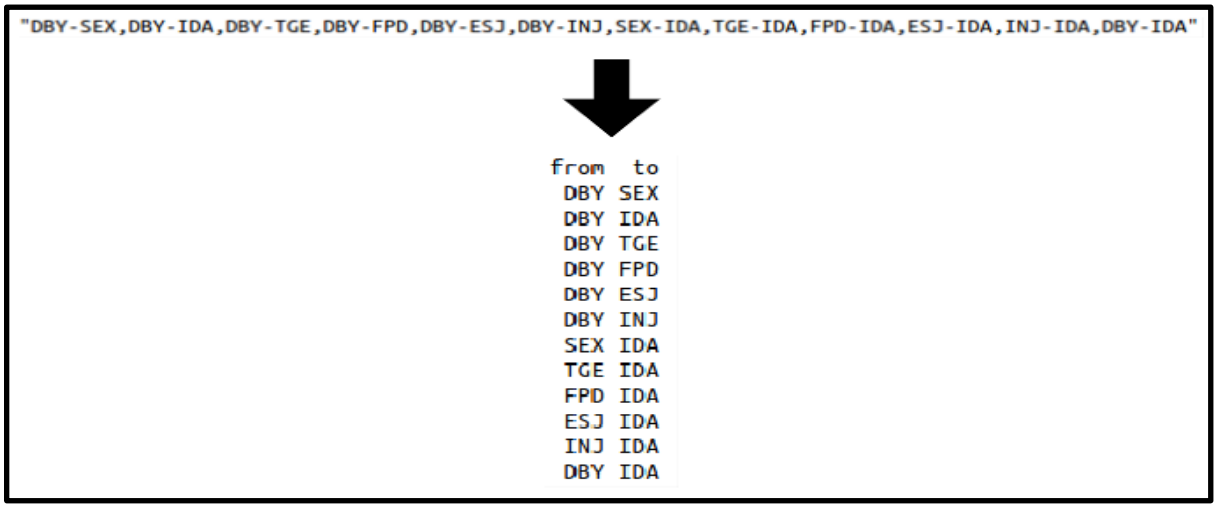
```
# Set the outcome var(s)
type.var <- "o" # setting to outcome.predictor.var function to set a black list for outcome
var.name <- "DBY" # setting this variable as a typically outcome
black.list <- bnpa::outcome.predictor.var(data.to.work, var.name, type.var, black.list)

# Set the predictor var(s)
type.var <- "p" # setting to outcome.predictor.var function to set a black list for outcome
var.name <- "IDA" # setting this variable as a typically outcome
black.list <- bnpa::outcome.predictor.var(data.to.work, var.name, type.var, black.list)

# Mount a white/black list in bnlearn sintax
black.list <- bnpa::mount.wl.bl.list(black.list)
```

Fonte: o autor, 2018.

Figura 31 - Um exemplo de lista negra, sintaxa do método bnpa seguida do resultado do pacote bnlearn



Fonte SCUTARI, 2014.

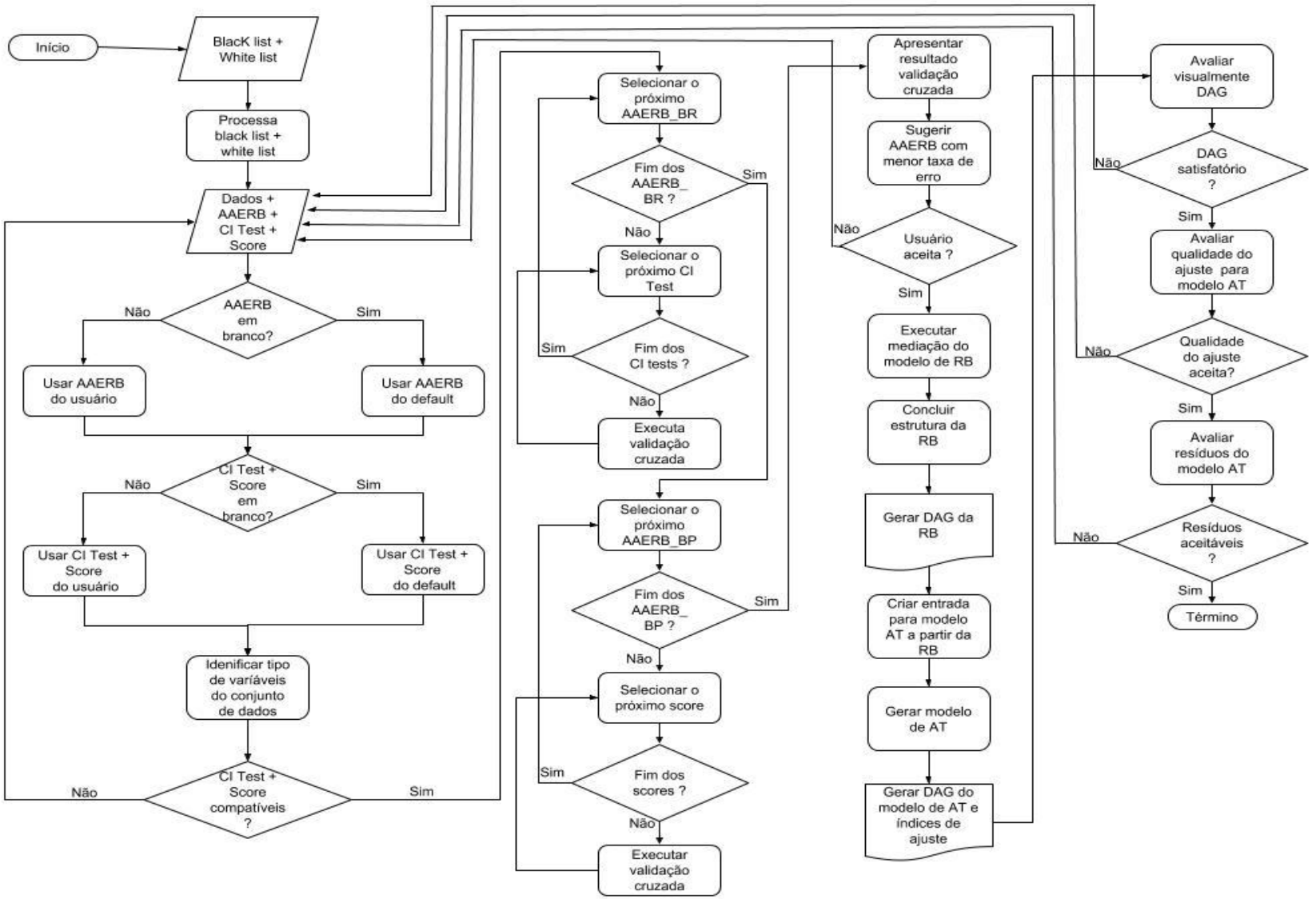
A pré determinação desses relacionamentos (conexões da RB) por meio da lista negra irá contribuir com a diminuição do tempo de processamento para aprender a RB e também com a diminuição de vieses durante a aprendizagem de estrutura da RB e da própria causalidade dos estudos epidemiológicos. O algoritmo 1 auxilia na criação automática de listas negras, no entanto, as listas negras também podem ser criadas manualmente. O pacote bnlearn também permite a criação de listas brancas, as quais indicam relacionamentos obrigatórios (o contrário das listas negras) e no caso deste estudo estas devem ser criadas manualmente.

Pré-verificação de parâmetros e compatibilidade com conjunto de dados

A processo de execução do método proposto por este estudo passa por várias etapas e verificações, as quais são representadas pela Figura 32. Na primeira etapa, o método recebe como parâmetros as listas brancas e negras e as processa. Este processamento consistem em converter as listas para o formato exigido pelo pacote bnlearn (SCUTARI, 2014). Durante esta mesma fase também é recebido como parâmetro o conjunto de dados, uma lista com AAERBs, uma lista com testes de IC que serão utilizados durante a aprendizagem de estrutura de RBs por meio de AAERBs baseados em restrições e uma lista de escores a serem utilizados por AAERBs baseados em pontuação. Recebido esses parâmetros o método verifica primeiro se a lista de AAERBs está vazia. Em caso negativo a lista será mantida com os AAERBs especificados pelo pesquisador, caso contrário, serão utilizados os algoritmos padrões do método que são quatro AAERB baseados em restrição (GS, IAMB, FAST.IAMB, INTER.IAMB) e dois AAERB baseados em pontuação (HC e TABU) implementados no pacote bnlearn. Em seguida se verifica se a lista de testes de IC e escore estão vazias. Em caso negativo se usa o que o pesquisador especificou, caso contrário se utiliza a lista de parâmetros padrões do método que são os mesmos especificados na seção 4.3.2.1 e dependem do tipo de variável.

Neste contexto implementou-se um processo no método que identifica se a variável é categórica ou numérica. Se categórica a função ainda identifica se é nominal, ordinal ou ainda dicotômica e mostra para o usuário para que este confirme os tipos identificados. Ainda se for categórica ordinal, o método mostra a classificação do fatores para que o usuário identifique se há algum erro ou não e confirme para o processo prosseguir. Se a variável for numérica, esta função ainda identifica se ela é discreta ou contínua. A título de exemplo se supõe que o método identifique as suas variáveis como categóricas ordinais, neste caso o teste de independência condicional disponibilizado para algoritmos baseados em restrição será Jonckheere-Terpstra (JT) e para algoritmos baseados em pontuação será log de probabilidade multinomial, critério de informação Akaike, critério de informação Bayesiana e pontuação equivalente à densidade posterior de Dirichlet. Após identificar o tipo de variável, o método verifica a compatibilidade do tipo com os testes de IC e escores. Caso não sejam compatíveis deve-se reespecificar o conjunto de dados, os testes de IC e escores novamente e reiniciar todo o processo. Caso sejam compatíveis se inicia o processo de aprendizagem da RB.

Figura 32 - Fluxo de funcionamento do processo de aprendizagem de estrutura de RB



Fonte: o autor: 2018.

A aprendizagem e validação da estrutura das RBs

Estando as listas brancas e negras prontas, o conjunto de dados pré-processados e sendo os algoritmos, os testes e escores compatíveis, inicia-se o processo de aprendizagem da estrutura da RB. Para isso o método percorre a lista de AAERBs primeiro os baseados em restrição e também percorre a lista de testes de IC. Para cada combinação "AAERB" + "Teste de IC" se executa uma validação cruzada. Em seguida o método percorre a lista de AAERBs baseados em pontuação e também a lista de escores. Neste caso, para cada combinação "AAERB" + "escore" também se executa uma validação cruzada. A validação cruzada tem como padrão 1000 rodadas por 10 conjuntos totalizando 10.000 estruturas de RB como resultado, no entanto este parâmetro também pode ser alterado pelo pesquisador e passado como parâmetro ao executar o método. Para que fique mais claro a Figura 33 apresenta a linha de comando e todos os parâmetros que o usuário pode passar para o método *bnpa*. Em ordem de apresentação os parâmetros são: lista branca, lista negra, conjunto de dados, lista de algoritmos baseados em restrição, lista de algoritmos baseado em pontuação, lista de testes de IC a serem usados pelos algoritmos baseados em restrição, lista de escores a serem utilizados pelos algoritmos baseados em pontuação, números de rodadas para validação cruzada e números de subconjuntos que serão gerados a partir do conjunto de dados.

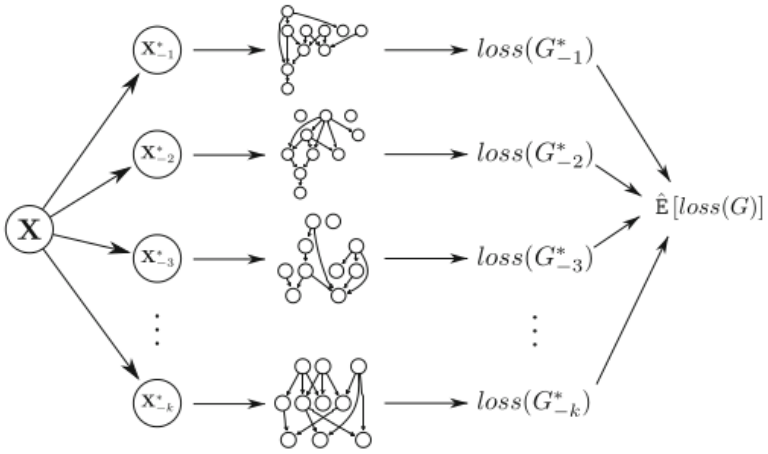
Figura 33 - Linha de comando para execução do método *bnpa* e todos os parâmetros que o pesquisador pode passar

```
gera.bn.cv.pa(white.list, black.list, data.to.work, cb.algorithms, sb.algorithms,
              cb.tests, sb.tests, number.of.runs, number.of.splits)
```

Fonte: o autor, 2018.

Para avaliar a capacidade de generalização da estrutura de RB aprendida a partir do conjunto de dados e evitar estimativas tendenciosas, utilizou-se o método de validação cruzada *k-fold* como recomendado por Koller e Friedman (2009) já implementado no pacote bnlearn (SCUTARI, 2014). Através deste procedimento, os dados foram divididos aleatoriamente em *k* subconjuntos, os quais foram utilizados, um por um, para validar os outros subconjuntos de dados *k*-1. Como o processo de validação cruzada é de natureza paralela (FIGURA 34), foi utilizado o pacote *paralell*, agora incorporado à linguagem R (TEAM, 2017), para paralelizar o processo de validação cruzada. Nossa estrutura usou *N*-1 números de núcleos de processador do computador disponível, como recomendado pelos autores do pacote *paralell*.

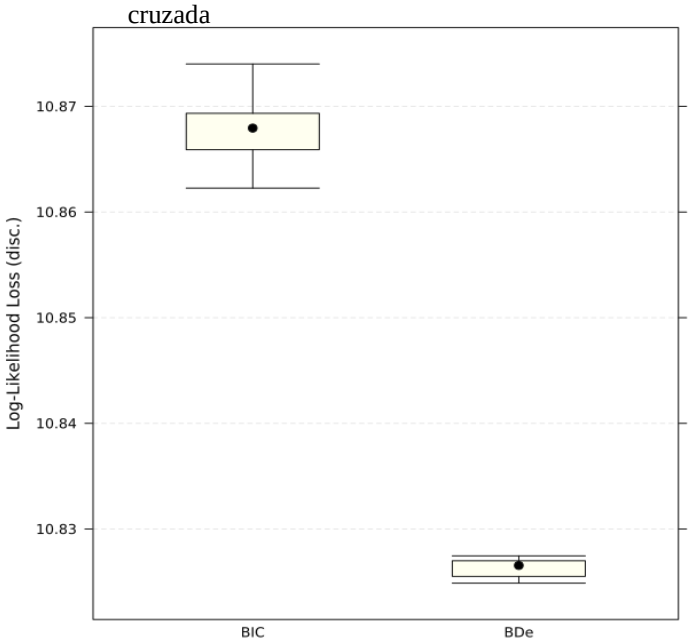
Figura 34 - Estimativa de validação cruzada de *K-fold* de uma função de perda para um algoritmo de aprendizado de rede bayesiana



Fonte: NAGARAJAN; SCUTARI; LÈBRE, 2013.

A função de validação cruzada implementada no pacote bnlearn gera uma estimativa de perda para o algoritmo de aprendizado de rede Bayesiana. Dessa forma, a estimativa de perda para cada algoritmo utilizado durante o processo de validação cruzada pode ser representado por um gráfico de caixas (*boxplot*) (FIGURA 35) facilitando a análise de desempenho de cada um deles.

Figura 35 - Diagrama de caixas (*boxplot*) representando a taxa de erro gerada durante o processo de validação cruzada



Fonte: Disponível em<http://www.bnlearn.com/examples/xval/>.

O método proposto *bnpa* coletará dados da validação cruzada executada pela combinação AAERB + testes de IC e/ou escores e ao final apresentará uma tabela e um *boxplot* com os resultados. Neste caso, o método recomendará a combinação que apresentar a

menor taxa de erro, no entanto, fica a cargo do pesquisador analisar esses dados e aceitar a sugestão do método ou não e até mesmo utilizar todas as estruturas geradas para somente no final do processo escolher o modelo definitivo.

Após a execução do processo de validação cruzada o método realiza automaticamente o processo conhecido como "mediação do modelo" (*model averaging*) (CLAESKENS, 2008). Este processo visa selecionar todos os arcos significativos das estruturas de RBs aprendidas com base em critérios recomendados pela literatura (NAGARAJAN, 2013). Os parâmetros utilizados para seleção são a força da relação entre duas variáveis (*strength*) e a direção (*direction*). A literatura, como em Sachs (2005) diz que esse limiar deve ser maior que 0.85, indicando que, para um arco ser considerado significante este deve estar em 85% das estruturas de RBs aprendidas. No entanto, Scutari e Nagarajan (2011) provaram que o processo de mediação da RB considerando um limiar de força de 85% tem o mesmo resultado que um considerando 50%. Mediante este problema eles formularam um algoritmo para calcular o limiar significante que será diferente para cada conjunto de RBs. A solução para este problema é apresentada em Scutari e Nagarajan (2011), está implementada no pacote bnlearn e é utilizada pelo método *bnpa*. Com relação a direção, arcos com probabilidade de direção exatamente igual a 0,5 são considerado escore equivalente e sua direção não pode ser identificada, com um valor maior que 0,5 fornece suporte para confirmar sua direção e menor que 0.5 não tem força para determinar a direção. A Figura 36 apresenta uma lista com arcos da RB aprendida a serem removidos e mantidos (processo de mediação da rede).

Figura 36 - Lista de arcos da estrutura de RB aprendida a ser removida (em vermelho) e a ser mantida (em azul)

	from	to	strength	direction
1	A	B	1.000	0.500
2	A	C	0.000	0.000
3	A	D	1.000	1.000
4	A	E	0.000	0.000
5	A	F	0.000	0.000
6	B	A	1.000	0.500
7	B	C	0.000	0.000
8	B	D	0.000	0.000
9	B	E	1.000	0.995
10	B	F	0.010	0.500
11	C	A	0.000	0.000
12	C	B	0.000	0.000
13	C	D	1.000	1.000
…				
27	F	B	0.010	0.500
28	F	C	0.005	0.500
29	F	D	0.000	0.000
30	F	E	1.000	0.995

Fonte: o autor, 2018.

Como resultado do processo de mediação do modelo de RB gerado pelo método *bnpa* será concebido o modelo final da estrutura de RB aprendida a partir dos dados. Lembrando que é possível conceber uma única estrutura ou se optar por criar todas as estruturas geradas pela combinação AAERB + testes de IC e escores.

Após a aprendizagem e escolha da melhor da estrutura de RB aprendida a partir dos dados, deve se prosseguir e utilizá-la para criar o modelo de entrada para dar início ao processo de criação do modelo de AT. A seção a seguir descreve as etapas necessárias para criação e validação deste modelo.

4.2.3 Criação do modelo de análise de trilhas

Como já foi explicado em seção anterior modelos de AT não descobrem a relação causal, mas dada a sua robustez estatística, AT combina bem as informações quantitativas dadas pelos coeficientes para fornecer uma interpretação quantitativa adequada (WRIGHT, 1934). A primeira etapa para conduzir um modelo de AT é construir um modelo de entrada que represente os relacionamentos hipotéticos. O próximo passo é executar as análises estatísticas e, em seguida, construir o grafo AT como saída e gerar as medidas inferidas. Este grafo representará as relações entre as variáveis (BRYMAN, 1990). Todo o processo para construir um modelo de AT representando a relação causal entre as variáveis de um conjunto de dados é tipicamente feito por estatísticos e / ou especialistas em MEE, um recurso caro e demorado, nem sempre imediatamente disponível para todos os pesquisadores. Com base nessa afirmação, justifica-se o uso de algoritmos de AAERBs para construir a estrutura do modelo de entrada de AT.

Para construir um modelo AT, o método *bnpa* usou o pacote R *lavaan* (ROSSEEL, 2012). O pacote *lavaan* foi desenvolvido para fornecer recursos para estimar uma grande variedade de análises de modelos estatísticos multivariados, incluindo AT, análise fatorial confirmatória, MEE e modelos de curva de crescimento. Se o modelo de entrada de AT tiver variáveis categóricas dicotômicas exógenas (VIs), estas precisam ser recodificadas como *dummy* (0/1); se elas forem ordinais, precisam ser codificados para refletir sua ordem e então serão tratadas como qualquer outra covariável (numérica). Para variáveis endógenas (VDs), se estar forem variáveis categóricas dicotômicas ou ordinais, o argumento "*ordered*" do pacote *lavaan* deve ser usado. Por exemplo, de acordo com Rosseel (2014) se um modelo de AT possui quatro variáveis categóricas dicotômicas e/ou ordinais (*v1, v2, v3, v4*) a sintaxe para geração de um modelo de AT é a mesma para modelos de AFC apresentada pela Figura 37.

Após isso, o pacote *lavaan*, com base nos parâmetros passados escolherá automaticamente o estimador para calcular os parâmetros.

Figura 37 - Sintaxe para criação de modelos com variáveis categóricas dicotômicas/ordinais para análise fatorial confirmatória (mesma sintaxe para AT)

```
fit <- cfa(myModel, data = myData,
           ordered=c("item1","item2",
                     "item3","item4"))
```

Fonte:ROSSEEL, 2014.

Para gerar o modelo de entrada de entrada de AT foi desenvolvido o Algoritmo 2. Esse algoritmo recebe como entrada uma estrutura de RB aprendida na etapa anterior, o conjunto de dados e os nomes para salvar o diagrama e os parâmetros de AT.

Na primeira etapa, o modelo de entrada de AT é construído usando a estrutura de RB aprendida. É executada uma varredura em cada variável do conjunto de dados e dentro desta se faz uma varredura em cada variável da estrutura de RB gerada, a qual representa um nó da RB. O objetivo aqui é verificar para cada variável do conjunto de dados o seu correspondente na RB e então avaliar se este nó da RB tem pais. Em caso positivo se inicia a criação do modelo de entrada do modelo de AT, no caso a variável será considerada endógena e seus pais variáveis exógenas, pelo menos para aquela situação. Na segunda etapa, o algoritmo verifica se existem variáveis endógenas do modelo de entrada de AT que são categóricas dicotômicas ou categóricas ordinais, em seguida cria uma lista com essas variáveis e uma segunda lista com as outras variáveis (numéricas). Na terceira etapa, o algoritmo verifica se há variáveis endógenas a serem declaradas como categóricas ordinais, em caso positivo ajusta o comando para criar o modelo de AT para usar o argumento “ordered” do pacote lavaan, caso contrário, se ajusta o comando sem esse argumento. Na quinta etapa, o algoritmo calcula uma variedade de medidas de ajuste para avaliar a qualidade do ajuste global do modelo de AT, gera uma tabela de resíduos e por fim o grafo do modelo AT e seus parâmetros são exportados para análise do pesquisador.

Figura 38 - Algoritmo 2 para criação do modelo de entrada de AT com base na estrutura de RB criada, geração de índices e exportação do grafo de AT e seus parâmetros

Algoritmo 2: Construtor de modelos de PA a partir de esturutras de RBs

Entrada: *bn* (estutura de RB aprendida), *ds* (o conjunto de dados usado para aprender a estrutura de RB), *dn* (o nome do documento para salvar os parâmetros do modelo de AT), *gn* (nome para salvar o grafo de AT)
Saida: Parâmetros do modelos de AT, matriz de correlação residual, grafo do modelo de AT
Inicializar : *pa.input.model*, *ordered.to.declare*, *cat.to.transform.into.numeric*, *fitted.model*, *fitted.measures*, *residuals.correlation*

```
# Construir o modelo de entrada do grafo de AT
Para cada variável em ds      # Percorre todas as variáveis do conjunto de dados
    Para cada variável em bn # Percorre todos os nós da estrutura da RB
        Se existe um nó pai
            pa.input.model ← pa.input.model + variável atual do ds ~ pai da variável na BN
        Fim se
    Fim para
Fim para
# Descobrir variáveis categóricas dicotômicas e ordinais
 Para cada variável em ds # Percorre todas as variáveis do conjunto de dados
    Se a variável atual no ds não tem pais na estrutura da RB aprendida e é categorica dicotômica ou ordinal
        ordered.to.declare ← ordered.to.declare + variável atual no ds
    senão
        cat.to.tranform.into.numeric ← cat.to.tranform.into.numeric + variável atual no ds
    Fim se
 Fim para
# Criar o modelo de AT
 Se existe categóricas orfinais/dicotômicas em numéricas
    fitted.model ← fit the model (pa.input.model, ds, ordered option (ordered.to.declare))
 senão
    fitted.model ← fit the model (pa.input.model, ds)
 Fim se
# Extrair e exportar as medidas de ajuste, a matriz de correlação residual e o grafo de AT
fitted.measures ← extract the fit measures (fitted.model)
residuals.correlation ←extract the correlation residual matrix (fitted.model)
export PA model parameters (fitted.measures, dn )
export(residuals.correlation)
export PA model graph (fitted.model, fitted.measures, gn)
```

Fonte: o autor, 2018.

4.2.4 Validação do modelo de análise de trilhas

A avaliação do desempenho do modelo PA foi feita através do conjunto de estatísticas de ajuste recomendado por (MAROCO, 2010; BEAUJEAN, 2014 e KLINE, 2015), portanto se utilizou os seguintes índices com seus respectivos limites de corte para avaliá-los como ajuste ruim, moderado ou bom:

1. RMSEA - Avalia se um modelo especificado tem uma aproximação razoável sobre os dados. Escores inferiores a 0,08 indicam melhor ajuste;
2. CFI - Avalia a porcentagem da qualidade do ajuste incremental do modelo proposto em relação ao modelo basal, o pior modelo possível que assume zero

covariáveis entre variáveis endógenas.Valores próximos a 1,0 indicam melhor ajuste;

3. SRMR- É um índice de ajuste absoluto que é uma estatística de má qualidade de ajuste. É uma versão padronizada da raiz quadrada média residual (RMR), que é uma medida da média absoluta da covariância residual. O ajuste perfeito seria representado por SRMR = 0, e valores cada vez mais altos indicam mau ajuste;
4. RMR - É uma medida da raiz média residual de covariância absoluta, pontuações próximas a 0,0 indicam melhor ajuste;
5. GFI - Semelhante ao R^2 na regressão, compara o ajuste do modelo proposto a um modelo saturado que permite que todas as variáveis se enquadrem. Pontuações maiores que 0,9 indicam um melhor ajuste.

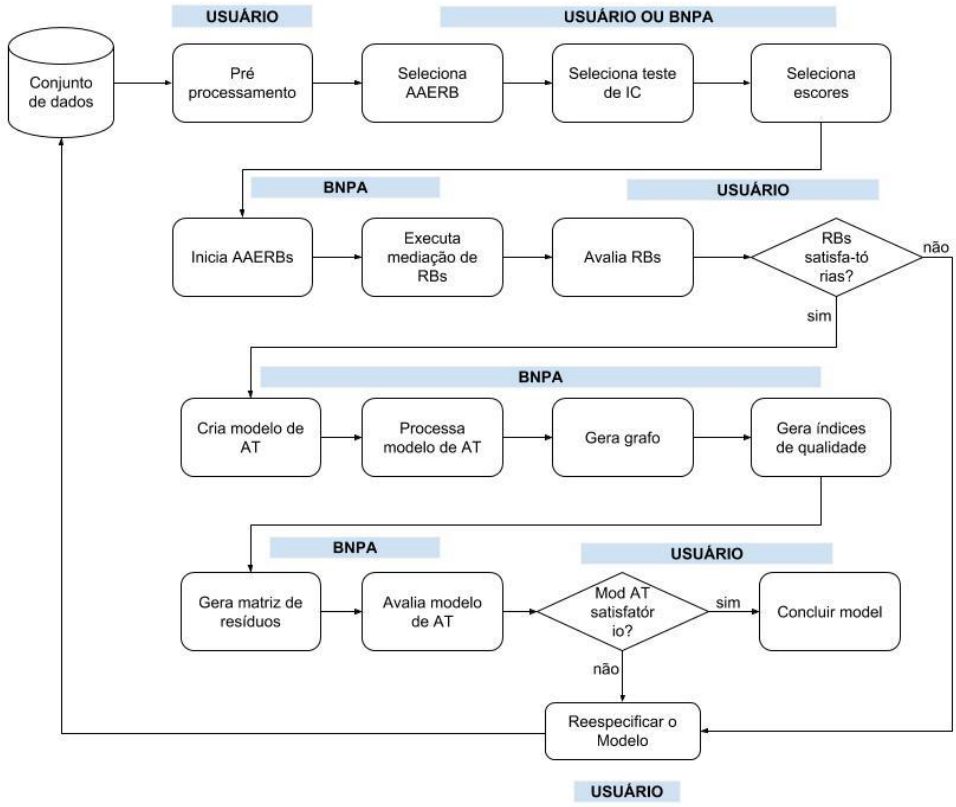
Utilizar somente índices de ajuste não implica necessariamente um bom modelo, portanto são necessários métodos adicionais. A diferença entre a correlação do modelo do pesquisador e a correlação do modelo de amostra é conhecida como correlação de resíduos. A literatura sobre MEE (BEAUJEAN, 2014; KLINE, 2015) afirma que os resíduos de correlação com valores absolutos superiores a 0,10 sugerem que o modelo do pesquisador não explica bem o modelo da amostra. Neste contexto se utilizou a matriz residual de correlação, gerada pelo pacote lavaan (ROSSEEL, 2014) para avaliar o ajuste do modelo.

Esclarecidos os métodos utilizados para elaboração deste estudo, a seção que segue a frente descreve como foram relizados os experimentos deste estudo.

4.3 EXPERIMENTOS

Para validar os resultados obtidos com o método proposto por esta tese foi necessário a realização de experimentos. Estes experimentos seguiram rigorosamente um protocolo para que os mesmos pudessem ser reproduzíveis em diferentes conjuntos de dados. É sabido que a falta de um procedimento ou a adição de um novo procedimento pode afetar o resultado final dos experimentos. Dessa forma, se apresenta seguir o protocolo dos experimentos realizados.A Figura 39 apresenta uma visão geral do protocolo que foi utilizado para a execução dos experimentos realizados.

Figura 39 - Protocolo de execução de experimentos



Fonte: o autor, 2018.

De um modo geral, embora o método *bnpa* forneça ferramentas para tal, o usuário se encarrega de fazer o pré-processamento dos dados, em seguida o usuário ou o método seleciona e define quais serão os algoritmos de aprendizagem de RBs e seus respectivos testes de IC e escores a serem utilizados durante o processo da aprendizagem das estruturas de RBS. A partir daí o *bnpa* assume o controle, inicia o processo de aprendizagem de estruturas das RBs e faz a sua mediação. Em seguida, o usuário inicia o processo de avaliação da(s) RB(s), se não for satisfatória este deverá especificar o modelo voltando ao conjunto de dados e seleção de AAERBs, testes de IC e escores novamente. Se a RB for satisfatória, o *bnpa* assume novamente o controle e cria o modelo de entrada de AT, processa o modelo, gera o grafo, os índices de qualidade do ajuste do modelo, a matriz de resíduos e uma tabela avaliando o modelo. Por fim o usuário analisa e decide se o modelo de AT é satisfatório ou não, em caso positivo se finaliza o processo e em caso negativo se inicia o processo de reespecificação do modelo.

4.4 CONSIDERAÇÕES FINAIS

Este capítulo apresentou o método proposto por esta tese, detalhando todas as etapas que a compoem e a metodologia adotada para execução dos experimentos. Foram apresentados detalhes sobre o método proposto e de como os experimentos forão realizados.

No próximo capítulo será apresentada uma análise detalhado dos resultados finais obtidos nestes experimentos.

5 ANÁLISE DOS RESULTADOS

5.1 CONSIDERAÇÕES INICIAIS

Este capítulo apresenta uma discusão detalhada sobre os resultados obtidos por meio dos experimentos realizados com o método *bnpa* desenvolvido nesta tese, cujo objetivo é capacitar modelos de AT a identificar causalidades por meio de AAERBs para criar modelos de entrada de AT e gerar automaticamente recursos para validação do modelo. Neste sentido, a seção 5.2 apresenta as características da base de dados utilizada para testes, a seção 5.3 justifica o uso do dados para realização de experimentos, a seção 5.4 apresenta as características do conjunto de dados final, a seção 5.5 detalha a geração das listas brancas e negras, a seção 5.6 fornece uma visão acerca do resultado obtido com a validação cruzada, a seção 5.7 demonstra como foram analisadas e escolhidas as RBs geradas, a seção 5.8 detalha a criação do modelo de AT, a seção 5.9 apresenta a forma como foi avaliada a qualidade do modelo de AT gerado, a seção 5.10 apresenta os efeitos diretos e indiretos gerados e a seção 5.11 apresenta as considerações finais do capítulo.

5.2 BASES DE DADOS

Por este estudo ser dedicado à área da saúde, torna-se necessário que se utilize bases de dados que apresentam características da realidade desta área. Neste contexto, é comum encontrar bases de dados com variáveis quantitativas (contínuas ou discretas) e variáveis qualitativas (nominais e ordinais). Portanto, para este estudo foi proposto utilizar uma base de dados de disponibilidade pública, mas que já tivesse sido explorada em outros artigos da literatura. As seções seguintes apresentam as características dessas bases de dados

A próxima seção descreve a origem e as características do banco de dados CCHS, o qual foi utilizado para experimentação do método *bnpa*.

5.2.1 Canadian Community Health Survey - CCHS

O CCHS é resultado de uma estudo transversal promovida no Canadá pelo *Canadian Institute for Health Information, Statistics Canada,* e *Health Canada* (STATISTICS CANADA, 2017). Este estudo foi conduzido entre os anos de 2001 a 2007, por meio de uma pesquisa respondida por aproximadamente 130.000 respondentes a partir de 12 anos. O objetivo do estudo foi coletar dados a respeito de determinantes da saúde, status da saúde da

população e utilização do sistema de saúde do país. Após 2007 a pesquisa foi conduzida todos os anos para fins de atualização. Este banco de dados é composto por 1.381 variáveis qualitativas nominais e ordinais e está disponível para uso público e mediante assinatura de termo de compromisso. O CCHS tem sido amplamente utilizado em diversos estudos científicos, uma revisão da literatura de 2013 avaliou os métodos estatísticos utilizados sobre este banco de dados e identificou 4.811 referências.

Feita a descrição do banco de dados e apresentação de suas características, a próxima seção descreve como foram realizados os experimentos deste estudo.

5.3 EXPERIMENTOS COM DADOS DO CCHS

A doença cardiovascular é a causa número um de mortes no mundo segundo a Organização Mundial da Saúde (OMS) sendo responsável por 17,7 milhões de óbitos. A OMS também afirma que é possível prever uma grande proporção de DCVs por meio de métodos de IA e MD que permitam a avaliação de fatores de risco, gerando suporte a aconselhamentos, tratamento e medicação mais adequados. Por esse motivo se decidiu utilizar o conjunto de dados CCHS, que tem informações sobre o paciente ter ou não DCV, em experimentos do método *bnpa* e verificar se este identifica características que influenciam de forma positiva ou negativa, direta ou indiretamente o risco de desenvolver DCV. Como primeiro passo para dar início ao estudo do conjunto de dados CCHS, um grupo de dois pesquisadores, doutores, com experiência em cardiologia e baseados na literatura selecionaram uma VD (CCC_121 ou HHD) e 13 VIs (QUADRO 4). Em seguida as variáveis foram renomeadas para que a estrutura dos grafos de RB e AT tivessem uma melhor apresentação. Foram removidos pacientes com menos de 18 anos, eliminadas respostas que não fossem de interesse e recodificadas variáveis categóricas dicotômicas conforme descrito na metodologia de pré-processamento. As variáveis categóricas identificadas não dicotômicas tiveram suas categorias ordenadas e verificadas. Ao final restarem 63.884 registros.

Quadro 4 - Variáveis selecionadas para este estudo

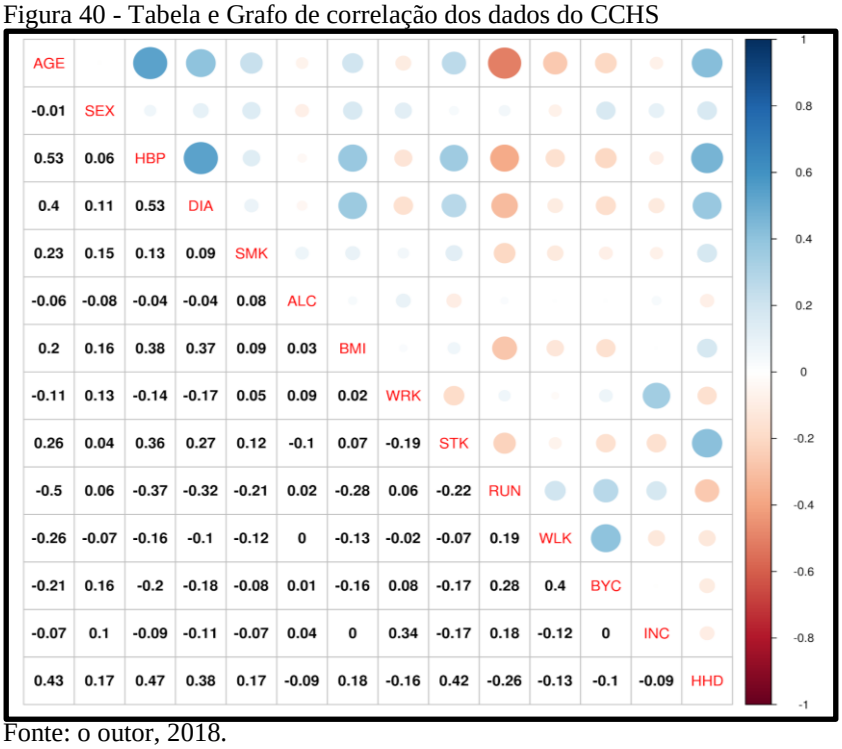
Nome original	**Novo Nome**	**Descrição**
DHHGAGE	AGE	Idade de 18 a 80 ou mais anos com 14 categorias
DHH_SEX	SEX	Sexo: 0 = Feminino, 1 = Masculino
CCC_071	HBP	Tem pressão alta: 0=Não, 1=Sim

Nome original	Novo Nome	Descrição
CCC_101	DIA	Tem diabetes: 0=Não, 1=Sim
SMKDSTY	SMK	É fumante: 0=Não, 1=Sim
ALCDTTM	ALC	Usa alcool: 0=Não bebe, 1=Bebe regularmente, 2=Bebe ocasionalmente
HWTGISW	BMI	IMC: 1=Peso normal, 2=Sobrepeso, 3=Obeso
GEN_08	WRK	Possui trabalho: 0=Não, 1=Sim
CCC_151	STK	Sofre efeitos de um AVC: 0=Não, 1=Sim
PAC_1J	RUN	Praticou atividade de corrida nos últimos 3 meses: 0=Não, 1=Sim
PAC_7	WLK	Pratica caminhada para o trabalho ou escola: 0=Não, 1=Sim
PAC_8	BYC	Vai de bicicleta para o trabalho ou escola: 0=Não, 1=Sim
INCGHH	INC	Renda total: 1=Nenhuma ou <20K, 2=20–39K, 3=40–59K, 4=60–79K, 5=80k+
CCC_121	HHD	Têm doença cardíaca: 0=Não, 1=Sim

Fonte: o outor, 2018.

Nenhum dado faltante ou *outlier* foi identificado conforme verificado pela função check.na e *check.outliers* do método *bnpa*. A verificação de colinearidade foi realizada por meio do grafo de correlação (FIGURA 40) e este não mostrou sinais de colinearidade.

Figura 40 - Tabela e Grafo de correlação dos dados do CCHS



Fonte: o outor, 2018.

Concluída a etapa de pré-processamento, o próximo passo foi elaborar as listas brancas e negras, detalhes sobre essa etapa do estudo são apresentadas na seção a seguir.

5.5 DEFINIÇÃO DAS LISTAS BRANCAS E NEGRAS

Após todas as verificações um grupo formado por dois pesquisadores, doutores, com experiência em pesquisa clínica e cardiologia se reuniram para definir quais seriam os componentes da lista branca e lista negra. Foi decidido que inicialmente não seria utilizado lista branca para não forçar nenhum relacionamento, pois este recurso seria utilizado somente se necessário. No entanto, foi necessário definir a lista negra que seria resultado da seleção da variável tipicamente de desfecho (VD) e das variáveis tipicamente preditoras (VIs). Neste sentido a variável "*HHD*" foi selecionada como tipicamente de desfecho e as variáveis "*AGE*" e "*SEX*" foram selecionadas como tipicamente preditoras. A lista foi processada (função *outcome.predictor.var*) pelo método *bnpa* e o resultado é apresentado no Quadro 5.

Quadro 5 - Lista negra contendo as variáveis tipicamente preditoras e de desfecho

From	To	From	To
HHD	AGE	STK	AGE
HHD	SEX	RUN	AGE
HHD	HBP	WLK	AGE
HHD	DIA	BYC	AGE
HHD	SMK	INC	AGE
HHD	ALC	HHD	AGE
HHD	BMI	AGE	SEX
HHD	WRK	HBP	SEX
HHD	STK	DIA	SEX
HHD	RUN	SMK	SEX
HHD	WLK	ALC	SEX
HHD	BYC	BMI	SEX
HHD	INC	WRK	SEX
SEX	AGE	STK	SEX
HBP	AGE	RUN	SEX
DIA	AGE	WLK	SEX
SMK	AGE	BYC	SEX

From	To	From	To
ALC	AGE	INC	SEX
BMI	AGE	HHD	SEX
WRK	AGE		

Fonte: o autor, 2018.

Com o conjunto de dados preparado e as listas brancas e negras definidas se prosseguiu com o início do processo de aprendizagem da estrutura da RB a partir do conjunto de dados. O resultado deste processo é apresentado na seção seguinte.

5.6 RESULTADOS REFERENTE A APRENDIZAGEM DA ESTRUTURA DE REDE BAYESIANA A PARTIR DOS DADOS E DO PROCESSO DE VALIDAÇÃO CRUZADA

Para dar início ao processo de aprendizagem de estrutura de RB pelo procedimento implementado no método *bnpa* (*gera.bn.cv.pa*), além do conjunto de dados, lista branca e lista negra foram passados também os parâmetros: *number.of.runs* ou número de rodadas para validação cruzada = 1000, *number.of.splits* ou número de conjuntos de dados utilizados pela validação cruzada = 10, *cb.algorithms* (algoritmos baseados em restrição) = *gs, iamb, fast.iamb* e*inter.iamb, sb.algorithms* (algoritmos baseados em pontuação) = *hc* e *tabu*, *cb.tests* (testes para os algoritmos baseados em restrição) = *jt* e *sb.tests* (testes para os algoritmos baseados em pontuação) = *aic*, *bic e bde*. Lembrando que os parâmetros "*cb.tests*" e "*sb.tests*" podem ser gerados automaticamente pelo método *bnpa* de acordo com a combinação AAERB x tipo de variáveis do conjunto de dados. Feito isso se iniciou o processo de aprendizagem das esturuturas de RB e validação cruzada para todos os algoritmos e testes passados como parâmetros. O resultado desse processo é apresentado pelaTabela 3 onde a taxa de erro médio variou de 8.77 a 8.85.

Tabela 3 - Resultado da validação cruzada

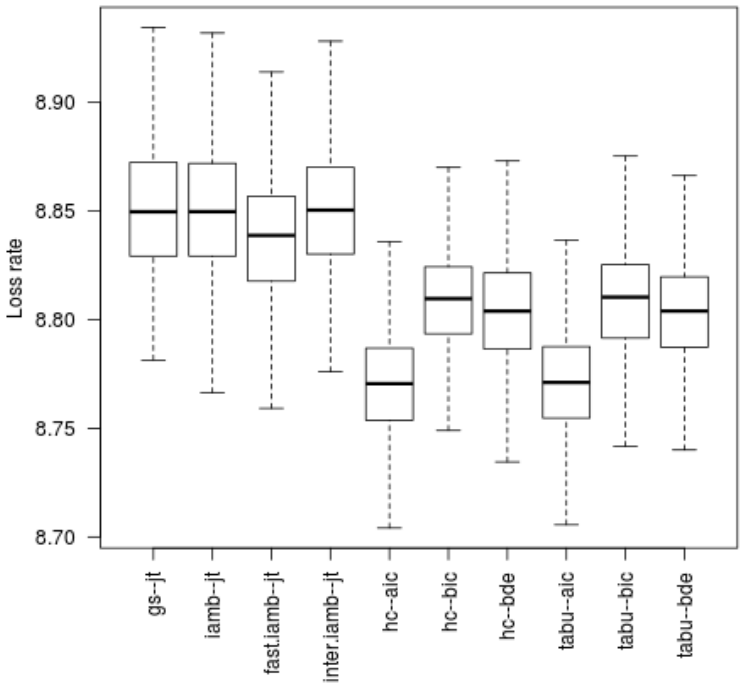
Algoritmo	Teste de IC / Escore	Taxa de Erro	Desvio Padrão
GS	JT	8,850	0,004
IAMB	JT	8,850	0,005
Fast IAMB	JT	8,838	0,003
Inter IAMB	JT	8,8507	0,005

HC	AIC	8,7716	0,001
HC	BIC	8,8091	0,001
HC	BDE	8,8059	0,001
Tabu	AIC	8,7717	0,001
Tabu	BIC	8,8096	0,001
Tabu	BDE	8,8046	0,001

Fonte: o autor, 2018.
JT - Jonckheere-Terpstra; AIC - Akaike information criterion score; BIC - Bayesian information criterion score; BDE - logarithm of the Bayesian Dirichlet equivalent score.

Além da Tabela 1 o método *bnpa* produz também um diagrama de caixas ou *boxplot* (FIGURA 41) o qual permite uma análise exploratória gráfica facilitando uma análise visual do desempenho dos AAERBs. De acordo com este diagrama, as caixas quase que se sobrepõem perfeitamente, neste caso todos os algoritmos têm aproximadamente o mesmo desempenho. Portanto a escolha da estrutura aprendida deverá ser com base na análise exploratória gráfica da mesma, o que acontecerá no próximo passo.

Figura 41 - Diagrama de caixa ou *boxplot* apresentando visualmente o resultado do processo de validação cruzada



Fonte: o autor, 2018.

O processo de validação cruzada gerou várias RBs compostas por vários relacionamentos entre as variáveis. Esses relacionamentos foram analisados e os relacionamentos considerados significantes foram selecionados pelo processo de mediação de rede conforme descrito na seção seguinte.

5.7 ANÁLISE E ESCOLHA DA ESTRUTURA DE RB APRENDIDA

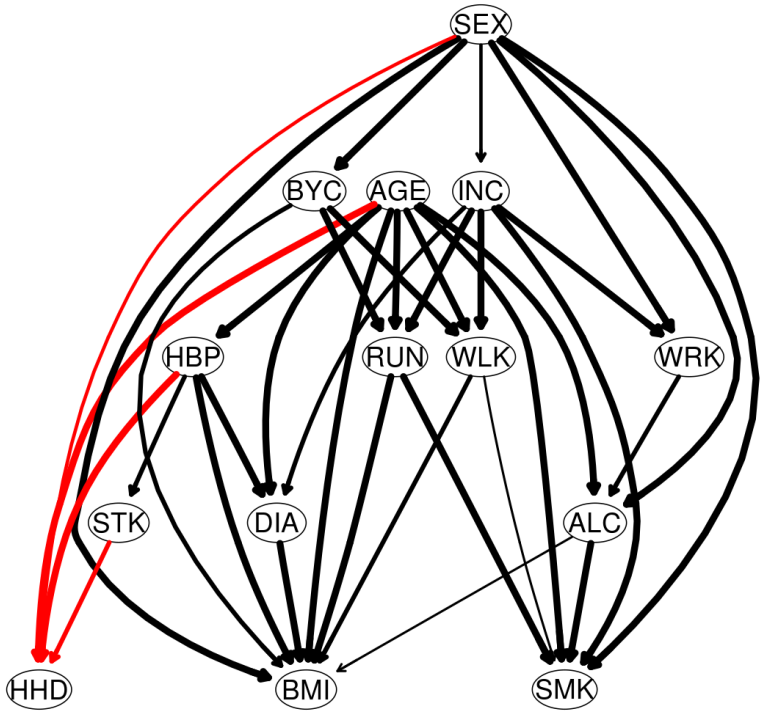
Após a execução do processo de validação cruzada foi executado o processo de mediação da rede (CLAESKENS, 2008). Durante este procedimento todos os arcos significantes da estrutura de RB foram selecionados, ou seja, aqueles relacionamentos com força igual ou maior ao limiar calculado pelo pacote bnlearn (SCUTARI, 2014) e com direção maior que 0,5 foram selecionados. Como resultado do processo se obteve 10 estruturas de RBs aprendidas automaticamente (FIGURA 42 a 52) que foram analisadas pelos pesquisadores, doutores, com experiência em pesquisa e cardiologia seguindo os seguintes critérios: a) DAGs apresentando o maior número de preditores corretos para a VD (*HHD*); b) a inexistência de relacionamentos incorretos entre as variáveis; c) a menor taxa de erro obtida durante o processo de validação cruzada.

A tabela 4 apresenta as características de todas as RBs resultantes do processo de mediação de rede. Nesta tabela se observa que número de arcos direcionados variou relavitamente, mas não excessivamente entre as categorias de algoritmos, exceto para tabu-aic e nenhum arco não direcionado foi gerado. O método *bnpa* também calculou 3 diferentes valores de escores (AIC, BIC e BDE) por meio do pacote bnlearn, os quais são utilizados para comparar modelos. Esses escores apresentaram valores semelhantes para todas as combinações AAERB x teste de IC ou método de pontuação, o que era esperado por esses algoritmos geralmente serem baseados na mesma família. Todas as estruturas aprendidas apresentaram conexões que corroboram com a literatura médica sobre DCV, no entanto a principal diferença ficou por conta do número de preditores que cada RB apresentou para a variável dependente HHD (tem doença cardiovascular). Os algoritmos baseados em restrição apresentaram mais preditores para a variável de desfecho HHD se comparados aos algoritmos baseados em pontuação. Por fim o algoritmo ‘fast.iamb’ apresentou o maior número de preditores para HHD (6) entre todos os algoritmos do experimento.

Tabela 4 - Resultado da mediação de RBs

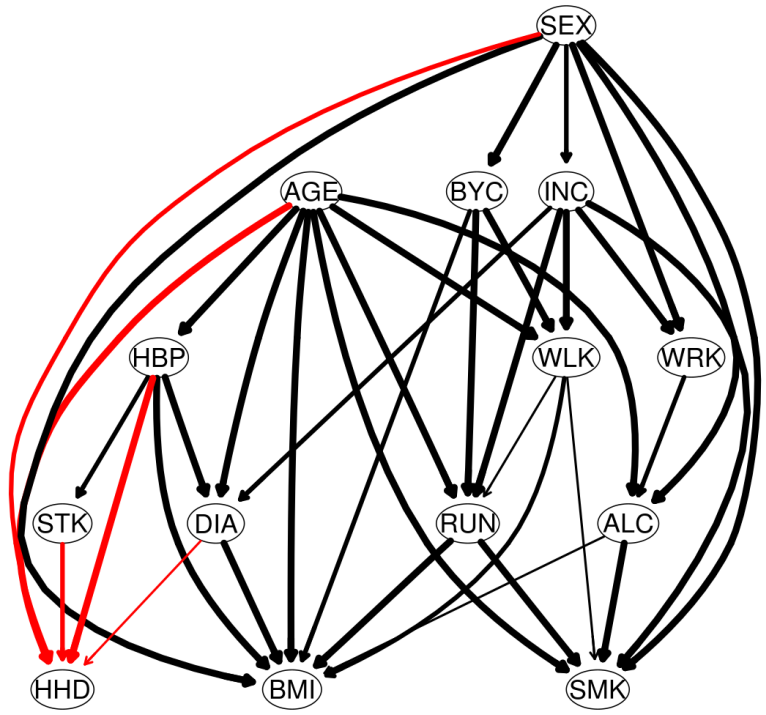
	Algoritmos baseados em restrição				**Algoritmos baseados em pontuação**					
	gs	**iamb**	**fast.iamb**	**inter.iamb**	**hc-aic**	**hc-bic**	**hc-bde**	**tabu-aic**	**tabu-bic**	**tabu-bde**
Arcos aprendidos	36	38	31	38	36	24	28	14	24	29
Direcionados	36	38	31	38	36	24	28	14	24	29
Não direcionados	0	0	0	0	0	0	0	0	0	0
Limiar de significância	0,48	0,50	0,41	0,49	0,51	0,54	0.50	0,52	0,56	0,52
Preditores para HHD	4	4	6	4	3	2	2	3	2	2
AIC	-178024.8	-178024.8	-178214.7	-178518.2	-177246.3	-177344.8	-177030.9	-177219.7	-177344.8	-177066.6
BIC	-188196.6	-188196.6	-186082.6	-194570.1	-188579.9	-178214.2	-178398.2	-188079.1	-178214.2	-178433.9
BDE	-184408.2	-184408.2	-183338.9	-189607.5	-186590.1	-177973.4	-177917.5	-185967.7	-177973.4	-177945.7

Figura 42 - Estrutura de RB gerada pelo AAERB *gs* em conjunto com o teste de IC *jt*



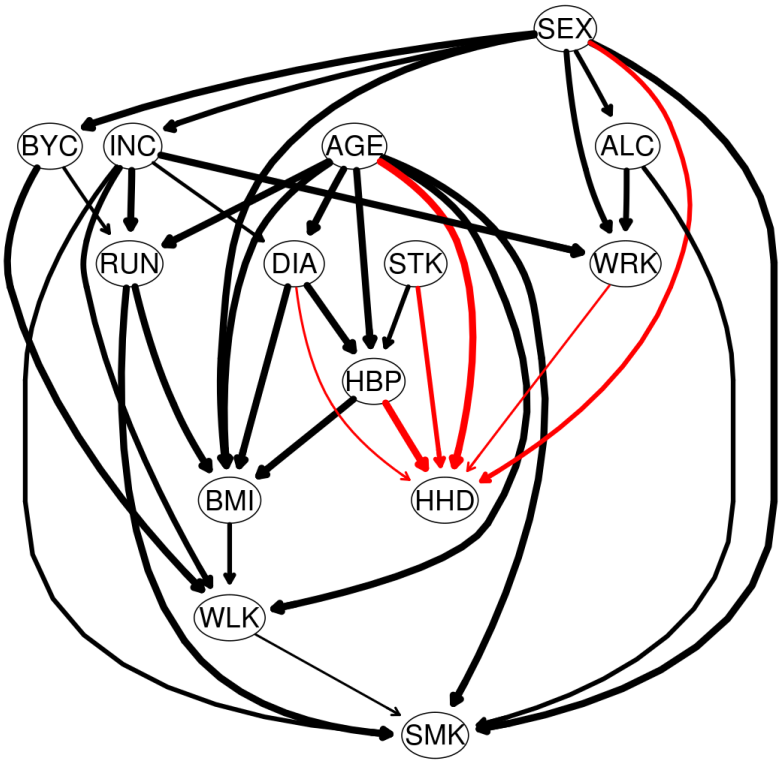
Fonte: o autor, 2018.

Figura 43 - Estrutura de RB gerada pelo AAERB *iamb* em conjunto com o teste de IC *jt*



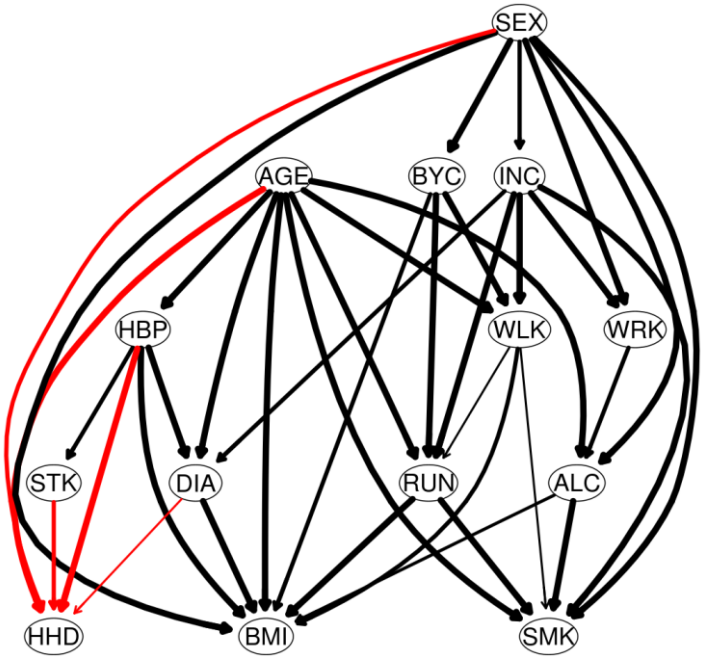
Fonte: o autor, 2018.

Figura 44 - Estrutura de RB gerada pelo AAERB *fast.iamb* em conjunto com o teste de IC *jt*



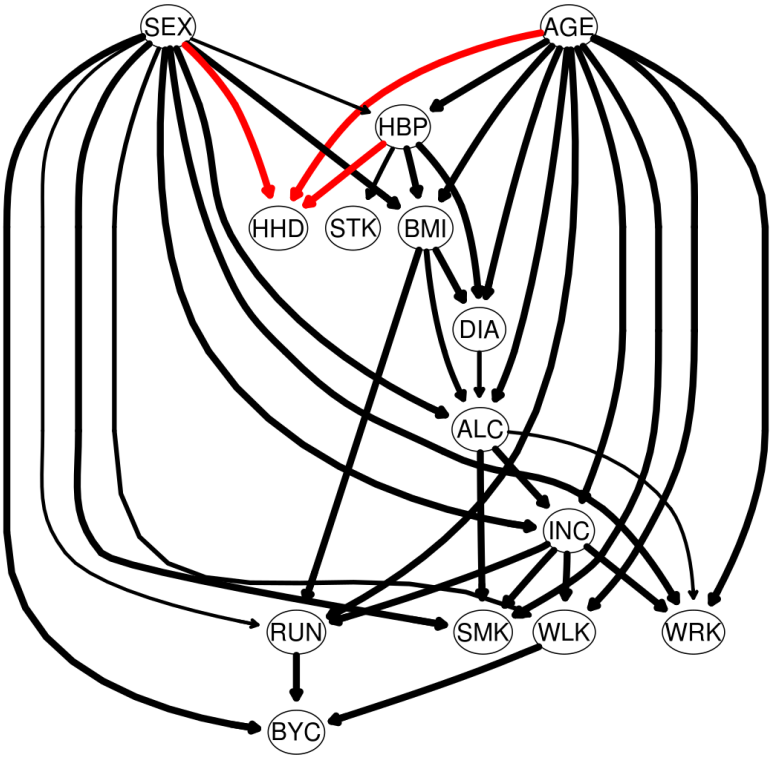
Fonte: o autor, 2018.

Figura 45 - Estrutura de RB gerada pelo AAERB *inter.iamb* em conjunto com o teste de IC



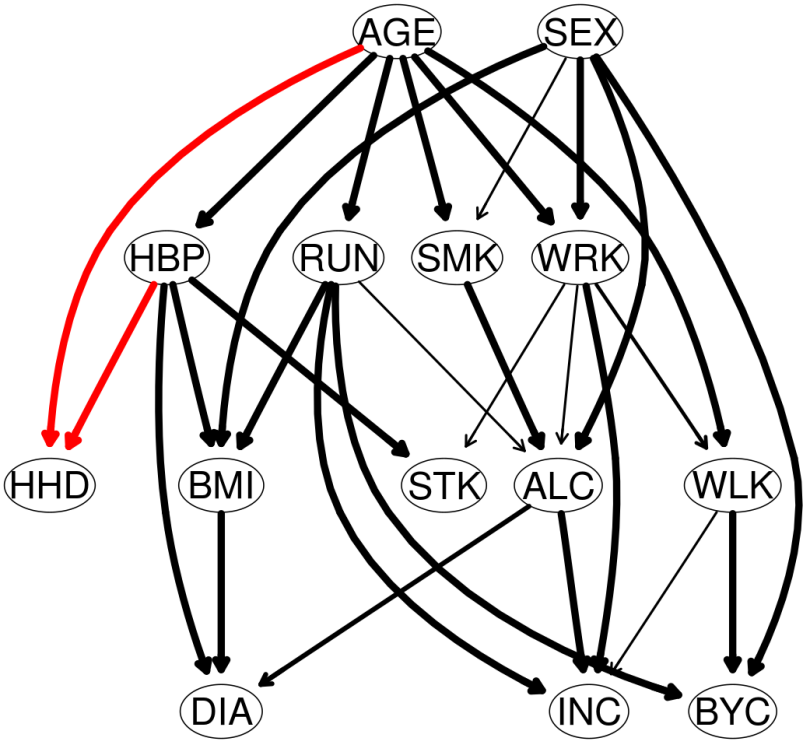
Fonte: o autor, 2018.

Figura 46 - Estrutura de RB gerada pelo AAERB *hc* em conjunto com o escore *aic*



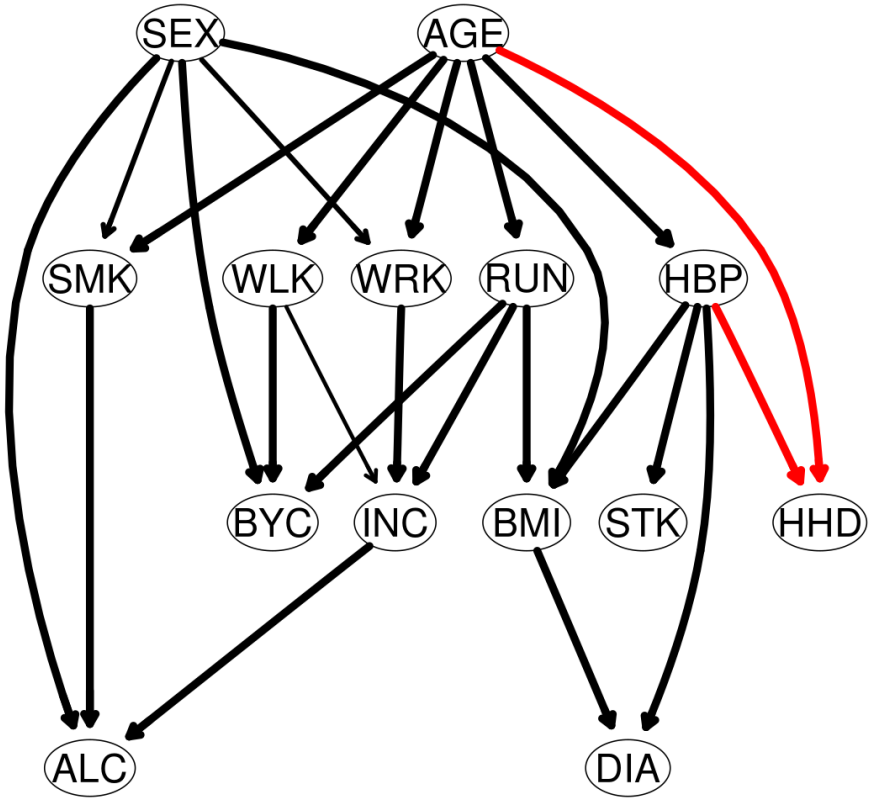
Fonte: o autor, 2018.

Figura 47 - Estrutura de RB gerada pelo AAERB hc em conjunto com o escore bde



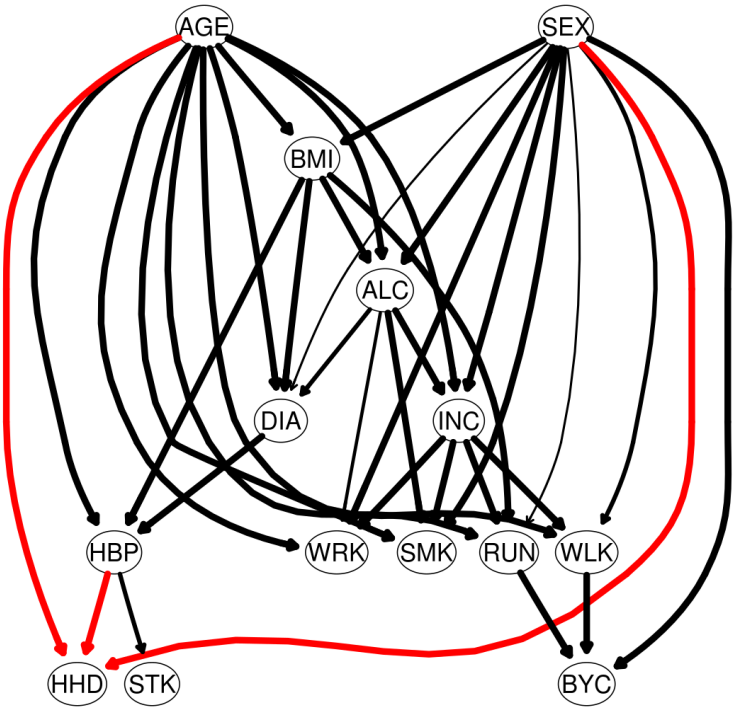
Fonte: o autor, 2018.

Figura 48 - Estrutura de RB gerada pelo AAERB hc em conjunto com o escore bic



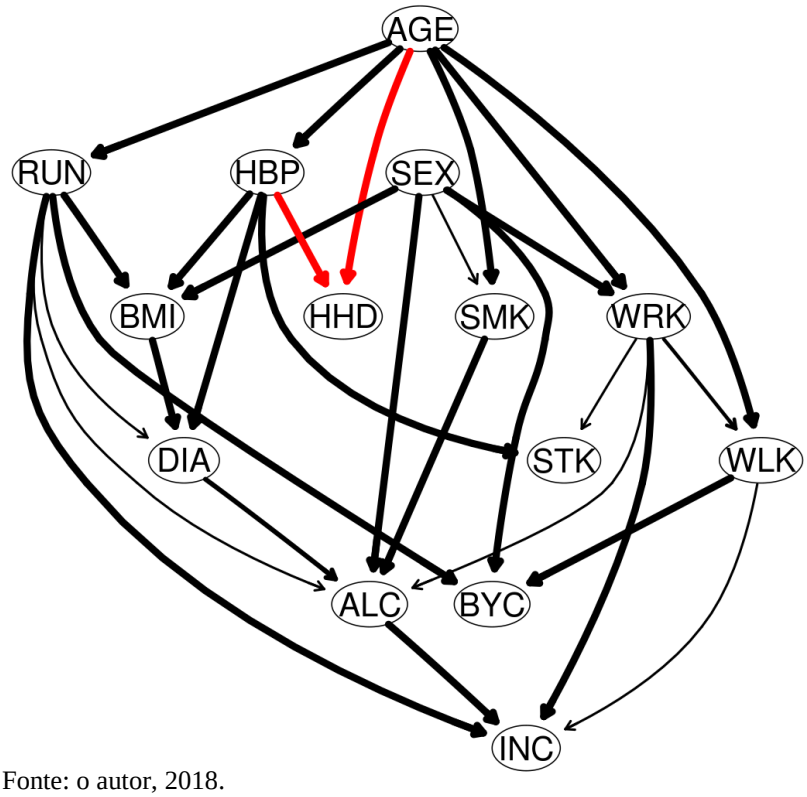
Fonte: o autor, 2018.

Figura 49 - Estrutura de RB gerada pelo AAERB tabu em conjunto com o escore aic



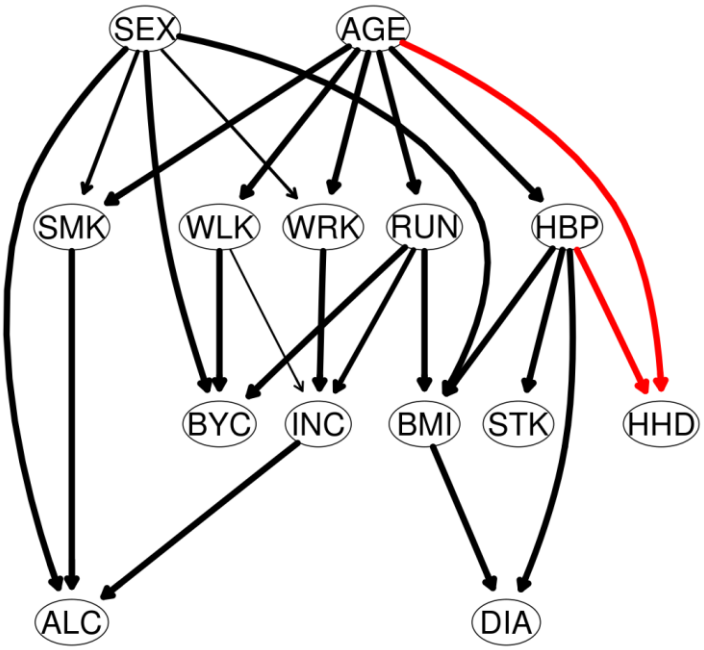
Fonte: o autor, 2018.

Figura 50 - Estrutura de RB gerada pelo AAERB tabu em conjunto com o escore bde



Fonte: o autor, 2018.

Figura 51 - Estrutura de RB gerada pelo AAERB tabu em conjunto com o escore bic



Fonte: o autor, 2018.

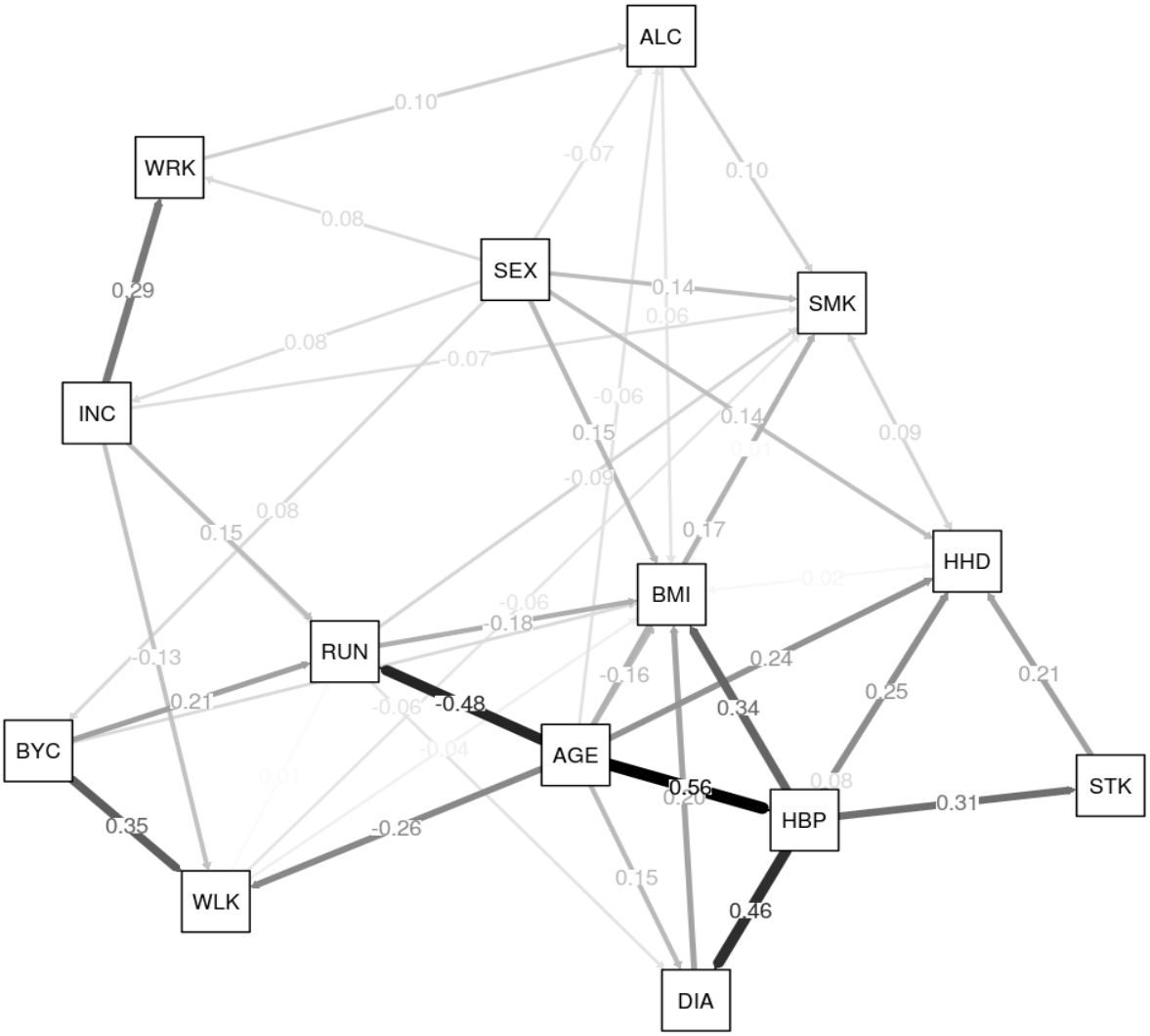
Como resultado final, a estrutura de RB gerada pelo algoritmo f*ast-iamb* foi escolhida (FIGURA 45) por atender os critérios estabelecidos pelos especialistas, principalmente por apresentar o maior numero de preditores para HHD de acordo com a literatura.

5.8 O PROCESSO DE CRIAÇÃO DO MODELO *DE* AT

Utilizando-se como base o algoritmo 2, foi implementado no método *bnpa* o procedimento (*gera.pa*) para criar, a partir da estrutura de RB aprendida o modelo de AT. Durante este procedimento se lê a estrutura da RB aprendida e monta a sintaxe para criação do modelo de entrada para AT, etapa esta criada sempre por especialistas em MEE, mas no caso do método *bnpa* o processo é executado automaticamente.

Este modelo é representado por linhas de comandos, mas na forma de arestas onde se tem o nó da esquerda mais o operador "~" e o nó ou nós da direita seguido por "+". O operador "~" representa uma regressão e a direção da borda é da direita para a esquerda. Esta sintaxe é descrita em detalhes em Rosseel (2014). Como exemplo, uma das linhas do modelo AT tem a seguinte aparência: HHD ~ SEX + AGE + STK + DIA + HBP + SMK + BMI. O gráfico do modelo de AT criado pela estrutura é apresentado na Figura 52. As arestas mais espessas indicam uma relação mais forte entre as variáveis.

Figura 52 - Modelo de AT gerado pelo método *bnpa* a partir da estrutura de RB aprendida a partir do conjunto de dados



Fonte: o autor, 2018.

Com o modelo de AT pronto, o pesquisador deve avaliar a qualidade do ajuste deste. Se o modelo tiver um bom ajuste deve ser aceito. A seção seguinte descreve esta avaliação.

5.9 A AVALIAÇÃO DE QUALIDADE DO AJUSTE DO MODELO DE AT AO CONJUNTO DE DADOS

Como visto na seção 2.5.5 existem diversos índices para avaliar a qualidade do modelo e seus respectivos valores de corte, os quais ditam se o ajuste tem desempenho ruim, moderado, bom ou muito bom. Além disso devido a grande quantidade de índices, optou-se

pela recomendação de (MAROCO, 2010) que avaliou diversas publicações sobre o assunto em conjunto com as opiniões de (BEAUJEAN, 2014 e KLINE (2015):

1. RMSEA - *Root Mean Square Error of Approximation;*
2. CFI - *Comparative Fit Index;*
3. SRMR - *Standardized Root Mean Square Residual.*
4. RMR - *Root Mean Square Residual;*
5. GFI - *Goodness-of- Fit Index;*
6. TLI - *Tucker-Lewis Index.*

Estes índices de qualidade do ajuste são apresentados na Tabela 5. Nesta tabela a primeira coluna representa os índices utilizados na avaliação de qualidade, a segunda o valor de corte estabelecido pela literatura, a terceira os valores gerados pelo método *bnpa* e as duas colunas subseqüentes apresentam valores dos índices de outros estudos similares obtidos durante a revisão da literatura (capítulo 3). Se utilizou apenas dois estudos pelos seguintes motivos: a) foram considerados apenas estudos usando os métodos de MEE e AT, excluindo-se estudos que utilizaram apenas RBs por estes apresentarem indices diferentes e b) a maioria dos estudos não reportou o valor do índice de qualidade utilizado. Em primeiro lugar, o resultado do método *bnpa* sinaliza um bom ajuste do modelo aos dados e em segundo lugar, a maioria de seus índices de qualidade do ajuste mostrou leve vantagem com relação aos mesmos índices das publicações.

Tabela 5 - Índices de ajuste do modelo AT ao dados

Índice	Valor de Corte	bnpa	De Heer et al (2012)	Vellone et al (2013)
RMSEA	<0,05	0,02	0,05	0,08
CFI	>0,90	0,95	0,96	0,98
SRMR	<0,05	0,05	ND	0,09
RMR	<0,05	0,04	ND	ND
GFI	>0,90	0,98	ND	ND
TLI	>0,90	0,92	0,95	ND

Fonte: o autor, 2018.

Como avaliação final do modelo de AT gerado, o método *bnpa* gerou por meio do pacote *lavaan* (ROSSEEL, 2014) a matriz de correlação residual, cujos valores absolutos não excederam a 0.10, exceto pelo relacionamento *INC* x *DIA*, confirmando um bom desempenho do modelo de AT gerado.

Figura 53 - Matriz de correlação do modelo de AT gerado

	HBP	DIA	SMK	ALC	BMI	WRK	RUN	WLK	BYC	INC	HHD
HBP	0,00										
DIA	-0,01	0,00									
SMK	0,04	-0,03	0,00								
ALC	0,01	-0,04	0,00	0,00							
BMI	0,00	0,05	0,02	0,04	0,00						
WRK	0,01	-0,04	0,09	0,05	0,09	0,00					
RUN	-0,06	-0,10	0,00	0,01	-0,03	0,02	0,00				
WLK	0,02	0,03	-0,08	-0,01	-0,03	0,02	0,00	0,00			
BYC	0,08	0,03	-0,04	-0,01	0,02	0,07	-0,02	0,06	0,00		
INC	-0,03	-0,12	-0,01	0,01	0,03	0,00	0,01	0,00	-0,01	0,00	
HHD	-0,01	0,05	0,09	-0,07	0,00	0,00	-0,03	0,00	-0,02	-0,06	0,00

Fonte: o autor, 2018.

5.10 OS EFEITOS DIRETOS E INDIRETOS GERADOS PELO MODELO AT

Uma vantagem em usar o método AT é a capacidade deste de apresentar os efeitos diretos e indiretos de uma variável para outra. A Tabela 5 apresenta os efeitos diretos e indiretos das variáveis sobre as possibilidades de um paciente ter doença cardíaca (*HHD*).

O efeito indireto é composto pela multiplicação do efeito da variável sobre outras variáveis que inflam a variável HHD e o efeito total a soma dessas duas medidas. Por exemplo, a variável “AGE” tem influência direta na variável “HHD“ de 0,24, mas tem também influência indireta de 0,14 que é resultado da influência que “AGE“ tem sobre “HBP“ de 0,56. Isso significa que “AGE“ também influencia indiretamente “HHD” por meio da variável “HBP”, neste caso o valor dessa influência será a multiplicação dos índices de cada trajetória que é igual a: 0,56 * 0,25 = 0,14. Portanto o efeito total de “AGE” sobre “HHD“ é 0,14 + 0,24 = 0,38.

Dessa forma, de acordo com a Tabela 6, idade influência mais o paciente a ter doencas cardiovasculares, seguida de pressáo alta e acidente vascular cerebral. Por fim o sexo também se demonstrou como uma variável importante como preditor de HHD.

Tabela 6 - Efeito direto e indireto obtido a partir do modelo de AT

Variável	Efeito		
	Indireto	**Direto**	**Total**
AGE	0,14	0.24	0,38
HBP	0.17	0.13	0.30
STK		0.21	0.21
SEX		014	0.14

Fonte: o autor, 2018.

5.11 CONSIDERAÇÕES FINAIS

Este capítulo apresentou uma análise detalhada dos resultados obtidos por meio dos experimentos executados para validação do método computacioal proposto pelo método *bnpa*. O próximo capítulo apresenta as principais conclusões deste estudo e sugestões para trabalhos futuros.

6 CONCLUSÃO

Esta tese apresentou o desenvolvimento do método *bnpa* como solução para a deficiência do método de AT em aprender um grafo causal a partir de dados. Este método ainda permite que o modelo seja "ajustado" pelos especialistas por meio das listas brancas e negras utilizadas em RBs. Outro benefício que o método *bnpa* fornece é a criação de modelos com um grande número de variáveis. Neste contexto, seria uma tarefa extremamente penosa e difícil senão dizer impossível para um *expert* devido a complexidade para definir o modelo. Neste cenário o conhecimento e/ou teoria se fazem muito mais necessários para modelar todas as possíveis relações entre as variáveis.

Os resultados apresentados pelos experimentos confirmaram a principal hipótese desta tese (*É possível construir modelos de AT com boa qualidade de ajuste a partir de um conjunto de dados por meio de AAERBs*) sugerem que técnicas de RBs podem ser um avanço no sentido de selecionar as variáveis mais relevantes para o modelo de AT, capacitando esta técnica com a habilidade de inferir causalidades. O índices de qualidade do ajuste juntamente com a matriz de resíduos sugerem que o modelo criado tem um bom ajuste.

Como limitação e/ou dificuldade destaca-se a falta de modelos de AT que pudessem ser reproduzíveis a partir de dados. Neste sentido, inicialmente foi idealizado utilizar o método proposto por esta tese para reproduzir um modelo final a partir de um conjunto de dados e comparar o resultado gerado pelo método *bnpa* e o modelo disponibilizado na literatura. No entanto, a maioria, senão todos os modelos de AT disponíveis em livros, tutoriais e aulas se baseiam em uma matriz de variância/covariância como entrada do método.

Para este estudo, utilizou-se apenas dados categóricos dicotômicos e ordinais. Para trabalhos futuros se pretende executar experimentos com dados contínuos e também com dados mistos (contínuos, categóricos dicotômicos e ordinais). Além disso, outros pacotes R que permitem a construção de RB serão incorporados no método *bnpa*.

REFERÊNCIAS

ACID, S. et al. A comparison of learning algorithms for Bayesian networks: a case study based on data from an emergency medical service. **Artificial intelligence in medicine**, v.30, n.3, p. 215-232, 2004.

AL-HAMADANI, B. An Emergency Unit Support System to Diagnose Chronic Heart Failure Embedded with SWRL and Bayesian Network. **International Journal of Advanced Computer Science and Applications**, v.7, n.7, p. 446-453, 2016.

ANDERS, R. L.; BALCÁZAR, H.; PAEZ, L. Hispanic community-based participatory research using a promotores de salud model. **Hispanic Health Care International**, v.4, n.2, p. 71-78, 2006.

ARBUCKLE, J. **Amos 17.0 user's guide**. SPSS Inc., 2008.

BARRETT, P. Structural equation modelling: Adjudging model fit. **Personality and Individual differences**, v.42, n.5, p. 815-824, 2007.

BEAUJEAN, A. A. **Latent variable modeling using R: A step-by-step guide.** Routledge, 2014.

BENTLER, P. M.; BONETT, Douglas G. Significance tests and goodness of fit in the analysis of covariance structures. **Psychological bulletin**, v.88, n.3, p. 588, 1980.

BENTLER, P. M. Comparative fit indexes in structural models. **Psychological bulletin**, v.107, n.2, p. 238, 1990.

BOLLEN, K. A. Sample size and Bentler and Bonett's nonnormed fit index. **Psychometrika**, v.51, n.3, p. 375-377, 1986.

BOLLEN, K. A.; STINE, R. A. Bootstrapping goodness-of-fit measures in structural equation models. **Sociological Methods & Research**, v.21, n.2, p. 205-229, 1992.

BOUCKAERT, R. R. **Redes de crenças bayesianas**: **da construção à inferência**. Tese (Doutorado) - Utrecht University, 1995.

BRYMAN, A.; CRAMER, D. **Quantitative data analysis for social scientists**. Taylor & Frances/Routledge, 1990.

BROADBENT, A. Causation and prediction in epidemiology: a guide to the "Methodological Revolution". **Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences**, v.54, p. 72-80, 2015.

BROWNE, M. W. et al. Alternative ways of assessing model fit. **Sage focus editions**, v.154, p. 136-136, 1993.

BUSSAB, W.; MORETTIN, P. **Estatística Básica**. 9.ed. São Paulo: Saraiva, 2017. 576p.

CASELLA, G.; BERGER, R. **Inferência estatística**. 2.ed. São Paulo: C. Learning, 2001. 588p.

CASTRO, M. A. et al. CO034. Overall and Central Obesity Indicators are Different Predictors of Metabolic Cardiovascular Disease Risk Factors: a Structural Equation Model Approach. **Archivos Latinoamericanos de Nutrición**, v 65, suplemento 2, 2015.

CHEN, W; SRINIVASAN, S. R.; BERENSON, G. S. Path analysis of metabolic syndrome components in black versus white children, adolescents, and adults: the Bogalusa Heart Study. **Annals of epidemiology**, v.18, n.2, p. 85-91, 2008.

CHENG, J.; BELL, D. A.; LIU, W. An algorithm for Bayesian belief network construction from data. In: **Proceedings of AI & STAT'97**, 1997. p. 83-90.

CHENG, J. et al. Learning Bayesian networks from data: an information-theory based approach. **Artificial intelligence**, v.137, n.1-2, p. 43-90, 2002.

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

CLAESKENS, G. et al. **Model selection and model averaging.** Cambridge Books, 2008.

COOPER, G. F.; HERSKOVITS, E. A Bayesian method for the induction of probabilistic networks from data. **Machine learning**, v.9, n.4, p. 309-347, 1992.

COSTA NETO, P. L. O. **Estatística**; 2.ed. São Paulo: E. Blücher, 2009. 280 p.

DALY, R.; SHEN, Q.; AITKEN, S. Learning Bayesian networks: approaches and issues. **The knowledge engineering review**, v.26, n.2, p. 99-157, 2011.

DE HEER, H. D. et al. A path analysis of a randomized promotora de salud cardiovascular disease–prevention trial among at-risk Hispanic adults. **Health Education & Behavior**, v.39, n.1, p. 77-86, 2012.

DE SÁ, A. G. C. **Evolução automática de algoritmos de redes Bayesianas de classificação**. 2014. Dissertação (Mestrado) - Universidade Federal de Minas Gerais, Minas Gerais.

DOS SANTOS, E. B. et al. Bayesian network classifiers: Beyond classification accuracy. **Intelligent Data Analysis**, v.15, n.3, p. 279-298, 2011.

FLORES, M. J. et al. Incorporating expert knowledge when learning Bayesian network structure: a medical case study. **Artificial intelligence in medicine**, v.53, n.3, p. 181-204, 2011.

FOX, J. **Polycor**:**polychoric and polyserial correlations**. R package version 0.7-8. Disponível em: <http://www.cran. r-project. org/web/packages/polycor/index. html>. Accesso em: 31 de Agosto de 2017.

FRIEDMAN, N.; GOLDSZMIDT, M.; WYNER, A. Data analysis with Bayesian networks: A bootstrap approach. In: **PROCEEDINGS OF THE FIFTEENTH CONFERENCE ON UNCERTAINTY IN ARTIFICIAL INTELLIGENCE**. Morgan Kaufmann Publishers Inc., 1999. p. 196-205.

GAMBORG, M. et al. Dynamic path analysis in life-course epidemiology. **American journal of epidemiology**, v. 173, n. 10, p. 1131-1139, 2011.

GATTI, E.; LUCIANI, D.; STELLA, F. A continuous time Bayesian network model for cardiogenic heart failure. **Flexible Services and Manufacturing Journal**, v.24, n.4, p. 496-515, 2012.

GLYMOUR, M. M.; KUBZANSKY, L. D. Causal inference in psychosocial epidemiology. **The Routledge International Handbook of Psychosocial Epidemiology**. Routledge, 2017. p. 35-60.

GARSON, G. D. **Structural Equation Modeling**. Asheboro, NC: Statistical Associates Publishers, 2015.

GREENLAND, S.; PEARL, J.; ROBINS, J. M. Causal diagrams for epidemiologic research. **Epidemiology**, p. 37-48, 1999.

GUJARATI, D. N.; PORTER, D. C. **Econometria Básica-5**. Amgh Editora, 2011.

HAIR JR., J.F. et al. **Multivariate data analysis**. Upper Saddle River, NJ: Prentice hall, 1998.

HAIR JR., J. F. et al. **Análise multivariada de dados**. Bookman, 2005.

HECKERMAN, D. Bayesian networks for data mining. **Data mining and knowledge discovery**, v.1, n.1, p. 79-119, 1997.

HERSKOVITS, E. H.; COOPER, G. F. Kutato: An entropy-driven system for construction of probabilistic expert systems from databases. **arXiv preprint arXiv:1304.1088**, 2013.

IRIONDO, J. M.; ALBERT, M. J.; ESCUDERO, Adrian. Structural equation modelling: an alternative for assessing causal relationships in threatened plant populations. **Biological Conservation**, v.113, n.3, p. 367-377, 2003.

JÖRESKOG, K. G.; SÖRBOM, D. **LISREL 7**: **A guide to the program and applications**. Spss, 1989.

KARCHER, C. **Redes Bayesianas aplicadas à análise do risco de crédito**. 2009. Tese (Doutorado) - Universidade de São Paulo, São Paulo.

KIM, Y. S. A path analysis model of health-related quality of life in patients with heart failure. **Journal of Korean Academy of Adult Nursing**, v.19, n.4, p. 547-555, 2007.

KLINE, R. B. Measurement models and confirmatory factor analysis. **Principles and practice of structural equation modeling**, v.22005, p. 133-145, 1998.

KLINE, R. B. Principles and practice of structural equation modeling. **Guilford publications**. 2015.

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

KORB, K. B.; NICHOLSON, A. E. **Bayesian artificial intelligence**. Florida: Chapman & Hall/CRC, 2003.

KORB, K. B.; NICHOLSON, A. E. **Bayesian Artificial Intelligence**, 2.ed. CRC Press, 2010.

KUH, Diana et al. Life course epidemiology. Journal **of Epidemiology & Community Health**, v.57, n 10, p. 778-783, 2003.

KUHNERT, P.; VENABLES, B.; ZOCCHI, S. S. **An introduction to R:software for statistical modelling & computing**. USP/ESALQ/LCE, 2005.

LARRANAGA, P. et al. Learning Bayesian networks by genetic algorithms: a case study in the prediction of survival in malignant skin melanoma. In: **ARTIFICIAL INTELLIGENCE IN MEDICINE IN EUROPE**. Springer, Berlin, Heidelberg, 1997. p. 261-272.

LAWLOR, D. A.; SMITH, G. D.; EBRAHIM, S. Commentary: The hormone replacement–coronary heart disease conundrum: is this the death of observational epidemiology?. **International Journal of Epidemiology**, v.33, n.3, p. 464-467, 2004.

MACCALLUM, R. C.; AUSTIN, J. T. Applications of structural equation modeling in psychological research. **Annual review of psychology**, v.51, n.1, p. 201-226, 2000.

MARGARITIS, D. **Learning Bayesian network model structure from data**. CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE, 2003.

MCINTOSH, C. N. Rethinking fit assessment in structural equation modelling: A commentary and elaboration on Barrett. **Personality and Individual Differences**, v.42, n.5, p. 859-867, 2007.

MAROCO, J. **Análise de equações estruturais:** Fundamentos teóricos, software & aplicações. ReportNumber, Lda, 2010.

MULAIK, S. A. et al. Evaluation of goodness-of-fit indices for structural equation models. **Psychological bulletin**, v.105, n.3, p. 430, 1989.

MULAIK, S. There is a place for approximate fit in structural equation modelling. **Personality and Individual Differences**, v.42, n.5, p. 883-891, 2007.

NAGARAJAN, R.; SCUTARI, M.o; LÈBRE, S. Bayesian networks in R. **Springer**, v. 122, p. 125-127, 2013.

NORSYS. **Nética Applications Help**. Disponível em:<www em: http://www.norsys.com/WebHelp/NETICA.htm>. Acesso em: dez de 2017.

OLIVEIRA, L. S. C.; ANDREAO, R. V.; SARCINELLI FILHO, M. Bayesian Network with Decision Threshold for Heart Beat Classification. **IEEE Latin America Transactions**, v.14, n.3, p. 1103-1108, 2016.

ORPHANOU, K.; STASSOPOULOU, A.; KERAVNOU, E. DBN-extended: a dynamic Bayesian network model extended with temporal abstractions for coronary heart disease prognosis. **IEEE journal of biomedical and health informatics**, v.20, n.3, p. 944-952, 2016.

OSBORNE, J.; WATERS, E. Four assumptions of multiple regression that researchers should always test. **Practical Assessment, Research e Evaluation**, Washington, v.8, n.2, p. 1-5, 2002.

PEARL, J. Causal diagrams for empirical research. **Biometrika**, v.82, p. 669-710, 1995.

PEARL, J. Causality: models, reasoning, and inference. **Econometric Theory**, v.19, n.675-685, p. 46, 2003.

PEARL, J.; **Causality**. Cambridge university press, 2009. 459 p.

PEARL, J.; GLYMOUR, M.; JEWELL, N. P. **Causal inference in statistics: a primer**. John Wiley & Sons, 2016. 135 P.

PEDHAZUR, E. J. **Multiple regression in behavioral research: Explanation and prediction**. Harcourt Brace Jovanovish College Publishers, 1997, 1058 p.

PENG, H.; DING, C. Structure search and stability enhancement of Bayesian networks. In: **Data Mining, 2003. ICDM 2003. Third IEEE International Conference on**. IEEE, 2003. p. 621-624.

QUINN, G. E. et al. Myopia and ambient lighting at night. **Nature**, v.399, n.6732, p. 113-114, 1999.

REBANE, G.; PEARL, J. The recovery of causal poly-trees from statistical data. **arXiv preprint arXiv:1304.2736**, 2013.

RICHMOND, R. C. et al. Approaches for drawing causal inferences from epidemiological birth cohorts: a review. **Early human development**, v.90, n.11, p. 769-780, 2014.

ROBINS, J. M.; HERNAN, M. A.; BRUMBACK, Babette. Marginal structural models and causal inference in epidemiology. **Epidemiology**, v.11, n.5, p. 550-560, 2000.
.

ROSSEEL, Y. Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). **Journal of statistical software**, v.48, n.2, p. 1-36, 2012.

ROSSEEL, Y. The lavaan tutorial. **Department of Data Analysis**: Ghent University, 2014.

SACHS, K. et al. Causal protein-signaling networks derived from multiparameter single-cell data. **Science**, v.308, n.5721, p. 523-529, 2005.

SATORRA, A.; BENTLER, P. M. A scaled difference chi-square test statistic for moment structure analysis. **Psychometrika**, v.66, n.4, p. 507-514, 2001.

SCHUMACKER, R. E.; LOMAX, R. G. **A beginner's guide to structural equation modeling**. Routledge, 2012.

SCUTARI, M. Learning Bayesian networks with the bnlearn R package. **arXiv preprint arXiv:0908.3817**, 2009.

SCUTARI, M.; DENIS, J. **Bayesian networks: with examples in R**. CRC press, 2014.

SCUTARI, M.; NAGARAJAN, R. On Identifying Significant Edges in Graphical Models of Molecular Networks. **arXiv preprint arXiv**:1104.0896, 2011.

SIMPSON, E. H. The interpretation of interaction in contingency tables. **Journal of the Royal Statistical Society**. Series B (Methodological), p. 238-241, 1951.

SINGH, M. et al. Building a Cardiovascular Disease Predictive Model using Structural Equation Model & Fuzzy Cognitive Map. In: **Fuzzy Systems (FUZZ-IEEE), 2016 IEEE International Conference on**. IEEE, 2016. p. 1377-1382.

SINGHY, M.; VALTORTA, M. Construction of Bayesian Network Structures from Data: a Brief Survey and an E cient Algorithm. **International journal of approximate reasoning**, v.11, p. 1-158, 1994.

SOUZA, T. V. **Aspectos estatísticos da análise de trilha (path analysis) aplicada em experimentos agrícolas**. 2013. Dissertação (Metrado) - Universidade Federal de Lavras.

SPIRTES, P.; GLYMOUR, C. N.; SCHEINES, R. **Causation, prediction, and search**. New York:Springer-Verlag,1993.

SPIRTES, P.; GLYMOUR, C. N.; SCHEINES, R. **Causation, prediction, and search**. MIT press, 2000.

STATISTICS CANADA. **Canadian community health survey.** 2011-2012. Health statistics division, statistics Canada, 2017.

SRINIVAS, S.; RUSSELL, S.; AGOGINO, A. Automated construction of sparse Bayesian networks from unstructured probabilistic models and domain information. In: **Machine Intelligence and Pattern Recognition**. North-Holland, 1990. p. 295-308.

STEIGER, J. H. Structural model evaluation and modification: an interval estimation approach. **Multivariate behavioral research**, v.25, n.2, p. 173-180, 1990.

STREINER, D. L. Finding our way: an introduction to path analysis. **The Canadian Journal of Psychiatry**, v.50, n.2, p. 115-122, 2005.

SU, C. et al. Using Bayesian networks to discover relations between genes, environment, and disease. **BioData mining**, v.6, n.1, p. 6, 2013.

SUZUKI, J. Learning Bayesian belief networks based on the MDL principle: An efficient algorithm using the branch and bound technique. **IEICE TRANSACTIONS on Information and Systems**, v.82, n.2, p. 356-367, 1999.

TEAM, R. Core. R: A language and environment for statistical computing. Vienna; 2014. **Austria: R Foundation for Statistical Computing Google Scholar**, 2017.

TSAMARDINOS, I. et al. Algorithms for Large Scale Markov Blanket Discovery. In: **FLAIRS conference**. 2003. p. 376-380.

TSHILIDZI, M. Causality, correlation and artificial intelligence for rational decision making. **World Scientific**, 2015.

VAN BUUREN, S. et al. Package 'mice'. **R package**. Disponível em: <http://cran. r-project. org/web/packages/mice/mice. pdf>. Accesso em: 31 de Agosto de 2017.

VANDENBROUCKE, J. P.; BROADBENT, A.; PEARCE, N. Causality and causal inference in epidemiology: the need for a pluralistic approach. **International journal of epidemiology**, v.45, n.6, p. 1776-1786, 2016.

VANDERWEELE, T. J.; VANSTEELANDT, S.; ROBINS, J. M. Marginal structural models for sufficient cause interactions. **American journal of epidemiology**, v.171, n.4, p. 506-514, 2010.

VELLONE, E. et al. Structural equation model testing the situation-specific theory of heart failure self-care. **Journal of Advanced Nursing**, v.69, n.11, p. 2481-2492, 2013.

VERMA, T. S.; PEARL, J. Equivalence and synthesis of causal models [Technical report R-150]. **Department of Computer Science, University of California, Los Angeles**, 1990.

VIGEN, T. **Spurious Correlations**. Hachette Books, 2015.

WEI, Z. et al. Using the Tabu-search-algorithm-based Bayesian network to analyze the risk factors of coronary heart diseases. **Zhonghua liu xing bing xue za zhi= Zhonghua liuxingbingxue zazhi**, v.37, n.6, p. 895-899, 2016.

WRIGHT, S. The method of path coefficients. **The annals of mathematical statistics**, v.5, n.3, p. 161-215, 1934.

YARAMAKALA, S.; MARGARITIS, D. Speculative Markov blanket discovery for optimal feature selection. In: **Data mining, fifth IEEE international conference on**. IEEE, 2005. p. 4.

YET, B. **Bayesian Networks for Evidence Based Clinical Decision Support**. 2013. Tese (Doutorado). Queen Mary University of London, 2013.

ZADNIK, K. et al. Vision: Myopia and ambient night-time lighting. **Nature**, v.404, n.6774, p. 143, 2000.

ZHANG, Z. Structural equation modeling in the context of clinical research. **Annals of translational medicine**, v.5, n.5, 2017.